# Expression and recognition of emotion in native and foreign speech : the case of Mandarin and Dutch

Zhu, Y.

# EXPRESSION AND RECOGNITION OF EMOTION IN NATIVE AND FOREIGN SPEECH

## THE CASE OF MANDARIN AND DUTCH

Published by
LOT
Trans 10
3512 JK Utrecht
The Netherlands

phone: +31 30 253 6111

e-mail: lot@uu.nl
http://www.lotschool.nl

Cover illustration: By the author.

ISBN: 978-94-6093-114-7

NUR 616

# EXPRESSION AND RECOGNITION OF EMOTION IN NATIVE AND FOREIGN SPEECH

# THE CASE OF MANDARIN AND DUTCH

PROEFSCHRIFT

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof. mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op donderdag 12 december 2013
klokke 16.15 uur

door

YINYIN ZHU

geboren te Beijing
in 1981

Promotiecommissie

| | |
|---|---|
| Promotor: | Prof. dr. Vincent J. van Heuven |
| Overige leden: | Prof. dr. Carlos Gussenhoven (Radboud University) |
| | Dr. Sylvie J. L. Mozziconacci |
| | Prof. dr. Marc van Oostendorp |
| | Prof. dr. Niels O. Schiller |

# Contents

# Acknowledgments

"*May he grant you your heart's desire, and fulfill all your plans. May we shout for joy over your victory, and in the name of our God set up our banners. May the Lord fulfill all your petitions.*" – ***Prayer for Victory***, (Psalms, 20:4-5)

First of all, I would like to take this opportunity to thank my supervisor, Vincent van Heuven, without whom I could never come this far. His patience, kindness, knowledge and encouragement were a great help to me during the PhD research.

Secondly, I thank the Leiden University Center for Linguistics and its Phonetics Laboratory, which provided me with supervision, technical support and recording facilities. I thank the Dutch advanced learners of Chinese and the lecturers from the Chinese Department at Leiden University who helped me with the experiments and the recordings. I thank the Dutch volunteers from all kinds of educational backgrounds who participated in the experiments. I thank the colleagues and the Chinese volunteers from the University of Science and Technology Beijing who contributed a big part in the whole research. I also thank those who assisted me in completing my dissertation in many other ways.

Finally, I thank my husband, Stephen Cooke, who has been giving me great mental and financial support. Without his support, I might not even have been able to start my PhD project three years ago. Also, I thank my families in China and the UK, without whom I would never be able to achieve my goal.

Leiden, May 2013

# Chapter One

# General Introduction

## 1.1 Introduction

In 1872 Charles Darwin published his book *The Expression of the Emotions in Man and Animals*, which has been highly influential for research on emotions (almost 3,000 citations according to the Institute for Scientific Information). However, Darwin himself did not define the term *emotion*. And in fact, the field of emotion research has found a consensual definition of this term elusive (Frijda 2000). According to the definition of Hess and Thibault (2009), emotions are considered to be relatively short-duration intentional states that entrain changes in motor behavior, physiological changes, and/or cognitions. Since Darwin started investigating emotions, there have been an increasing number of studies on perception and production of emotions through different channels, for example, through audio, visual or audio-visual sensory input. Studies on emotion were traditionally carried out in the fields of psychology, physiology, biology and were extended into other fields rapidly later. In the recent years, studies on vocally, facially or vocally-facially produced emotions have been conducted in the areas of sociology, linguistics, pathology, computer science, neuroscience, musicology and second language acquisition. In addition, there have been an increasing number of studies on perception or production of emotion cross-culturally and/or cross-linguistically. Previous studies have shed light on various aspects, for instance, why humans are able to perceive and produce emotions (Darwin 1872); through what cues humans perceive and produce emotions (Chang 1985, Chen 2005, Darwin 1872, Huttar 1968, Ohala 1984, Scherer 1979, etc.); how well humans can perceive vocal emotions in their native language, or in their second language or even in an unknown language (Chen 2005, Scherer et al. 1986, Van Bezooijen 1984, etc.); what the differences are between humans and machines in the perception and production of emotion (Ang et al. 2002, Bänziger et al. 2009, etc.); and what factors may limit the expression of emotion (Ross et al. 1986, etc.). Furthermore, research methods adopted in the previous studies were diverse, varying from traditional field work and experimental studies to meta-analysis based on literature review and existing corpus analysis.

It is worth briefly reviewing some important findings and conclusions of previous studies before starting a detailed literature review. It is claimed by some researchers that perception of emotion is universal. However, some other researchers believed that it is universal to some extent, but it is more likely to be culture-or-language specific or emotion-specific. Some studies argued that, in fact, perception of emotion combines both universal and culture-or-language specific cues. In addition, some previous studies found that perception of emotion through the audio-visual channel is more salient than

that is through the audio or visual channel only. Moreover, previous studies also showed that emotion is generally better recognized when expressed by a speaker of the same cultural group as the listeners. Previous studies also indicated that automatic recognition of human-produced emotions can reveal some of the acoustic cues that humans use to perceive and produce emotions. Apart from that, some studies in the field of neurolinguistics even showed the location in the hemisphere in which emotion is produced. Although previous findings of perception and production of emotion are abundant, there are still issues which have not been well investigated. For instance, previous studies did not give us a clear picture of how well listeners of a non-tonal language can perceive emotions produced in a tonal language (especially through the audio channel only), even though some of the previous studies had touched on this topic. Neither did previous studies give us any relatively clear views of how well L2 speakers of a language can vocally produce emotion in the L2 compared to native speakers, especially when the L2 is a tonal language but the L2 speakers' L1 is not. It is also not clear whether a speaker can vocally produce emotions in his L2 as well as he does in his native language. In other words, does L2 limit the expression of emotion to some extent?

The first aim of the present PhD study, therefore, is to use an experimental approach to investigate how well native and non-native listeners of a tonal language perceive vocal emotions portrayed in a tonal language. Non-native listeners in this dissertation will include naïve listeners and advanced L2 learners of the tonal language who share the same L1 as the naïve listeners. Secondly, I am going to investigate whether L2 speakers of a tonal language are able to vocally produce emotions in the L2 as well as they do in their L1; also, I will study how well native, naïve listeners and advanced L2 learners of a tonal language perceive vocal emotion expressed by L2 speakers of the tonal language. An acoustic analysis will be conducted thereafter to identify the vocal correlates that speakers and listeners use in the production and the perception of the vocal emotions. Finally, I will determine whether the 'in-group advantage' found by other researchers is universal, claiming that listeners generally better recognize emotional prosody produced in their L1 than in an unknown language.

A detailed literature review of what other researchers have done and the main findings of them will be provided in Chapter 2. In addition, in the same chapter there will be a description of the experimental design and the research methods that will be used in the later chapters.

## 1.2 Linguistic backgroud

### 1.2.1 Tonal language vs. non-tonal language

Tone is the linguistic use of pitch to distinguish meanings of words. It is used in many of the world's languages.[1] Tone is an abstract linguistic property. It is expressed mainly through vocal pitch, which in turn is determined mainly by the repetition rate of the vocal fold vibration. Therefore, tone is controlled by the larynx and possibly arose historically from the influence of laryngeal contrasts (such as voicing) in consonants. Languages may contrast up to four level tones, and maximally two different rises and/or falls. Typical tonal languages include most of the languages of sub-Saharan Africa, East and Southeast Asia, and Central America; many in North and South America and the Pacific; and even a number of languages of Western Europe, such as Swedish, Norwegian and even some varieties of Dutch/German (Yip 2006).

A tone language, then, is a language in which the pitch of the voice can change the meaning of the word. This is distinct from intonation, in which pitch changes may signal sentence-level meanings such as questions or surprise. A tonal language, therefore, is in contrast to a non-tonal language, which does not regularly use pitch change to distinguish lexical meaning, for example: English, German, French or Japanese.

#### 1.2.1.1 Chinese

Mandarin, or Standard Chinese, is a Sino-Tibetan tonal language which uses a wide pitch range (with pitch movements up to 12 semitones, Xu 1999). It has monosyllabic words and a simple syllable structure (Duanmu 2007a, b). Chinese is the first language of over 1 billion speakers. There are several dialect families of Chinese (each in turn consisting of many dialects), which are often mutually unintelligible (Cheng 1997, Tang 2009, Tang & Van Heuven 2009). However, there are systematic correspondences among the dialects and it is easy for speakers of one dialect to pick up another dialect rather quickly. The largest dialect family is the northern family (also called the Mandarin family), which comprises over 70% of all Chinese speakers. Standard Chinese (also called Mandarin Chinese) is a member of the northern family; it is based on the pronunciation of the Beijing dialect (Duanmu 2006). Mandarin Chinese has four tones: level, rising, falling-rising and falling (Chao 1948). The same segmental sequence may carry different meanings depending on the tone. For example, the meaning of Mandarin Chinese *ma* with Tone 1 is 'mother', the Tone 2 version means 'hemp', and the Tone 3 and 4 meanings are 'horse' and 'scold', respectively (e.g., Jongman et al. 2006). Mandarin Chinese is used in this dissertation to investigate how L1 and L2 speakers of a tonal language perceive vocally produced emotions in a tonal language.

[1] The World Atlas of Linguistic Structures (WALS, Comrie et al. 2005) lists 220 tone languages versus 307 no-tone languages (chapter 13); at the same time it lists 502 stress languages, divided in chapter 14 between 282 with fixed stress (281 in chapter 15) versus 220 with no fixed stress (219 in chapter 15). Van Zanten & Goedemans (2007: 64) estimate that languages with stress-based word prosody, tone-based systems and languages without word prosody occur in 80, 16 and 4% of the world's languages, respectively.

#### 1.2.1.2 Dutch

According to Nederlandse Taalunie (2005), Dutch is a West Germanic language which belongs to Indo-European languages and which is the native language of most of the population of the Netherlands. It is a non-tonal language, which contrasts with tonal languages, such as Mandarin, Thai and Vietnamese. Dutch is also spoken in other regions, such as the northern part of Belgium, Surinam, Aruba, Curaçao and Sint Maarten, and is closely related (and mutually intelligible to a considerable degree, Gooskens & Van Bezooijen 2006) with Afrikaans (spoken in South Africa). Moreover, Dutch is a stress-accent language, and has a rather restricted pitch range (De Pijper 1983, 't Hart et al. 1990), with often long, polysyllabic (compound) words that may contain complex consonant clusters (Booij 1995). Dutch has a quantity-sensitive stress system, which means that the heaviest syllable in the word – all else being equal – carries the main stress (Kager 1989, Langeweg 1988). For many speakers, their Dutch is coloured to some extent by the rural or urban dialect that they speak. However, only standard Dutch is used in this dissertation. Standard Dutch, however, has many regional varieties, which are reminiscent of (but very different from) the local dialects spoken in the area (Van Heuven & Van de Velde 2010).

### 1.2.2 Tone and emotional prosody

#### 1.2.2.1 Tone and Chinese lexical tones

In phonetics, tone is considered as a suprasegmental (or prosodic) phenomenon, which is predominantly expressed by vocal pitch. Specifically, tone is a feature of the lexicon, being described in terms of prescribed pitches for syllables or sequences of pitches for morphemes or words (Cruttenden 1986: 8); i.e. pitch distinguishes the meanings of words (Pike 1948: 3). The main acoustic correlate of tone (pitch) is the fundamental frequency of the speech signal, known as F0 – the number of times per second that the vocal folds complete a cycle of vibration. It ranges from a low of around 80 cycles per second (hertz or Hz) for the lowest speaking pitch of a male voice, to a high of around 400 cycles per second for the highest speaking pitch of a female voice. Generally, the (low) male and (high) female pitch ranges are distinct. As a result, the high tone of a male voice typically has an F0 that is lower than the low tone of a female or a child's voice (Yip 2006).

Previous phonetic studies have examined the fundamental frequency contours of Mandarin Chinese tones (e.g., Chuang et al. 1972, Dreher & Lee 1966, Dreher et al. 1969, Howie 1970, Liu 1924, Moore & Jongman 1997, Rumjancev 1972, Wang et al. 1967). These studies indicate that F0 height and F0 contour are the primary acoustic parameters to characterize Mandarin tones. In general, Tone 1 is high and relatively level over mostits duration. Tone 2 exhibits a rise for much of its duration, where the onset of the rise occurs in the middle region of the F0 range and ends at a point approaching the F0 height of Tone 1. The Tone 3 contour occupies the lowest region of the F0 range overall, although extending at least to the midpoint of the range by the offset. The Tone 3 onset is variable and can be close in frequency to that of Tone 2. Tone 4 begins high and falls to the bottom of the range (e.g. Jongman et al. 2006). The

pitch range of the four lexical tones of a male Chinese speaker extends normally from 80 to 223 Hz; and the one of a female Chinese speaker is generally from 165 to 352 Hz (Wu 1986).

#### 1.2.2.2 Emotional prosody

In order to know what emotional prosody is, it is helpful to first understand what prosody is. Prosody literally means 'accompaniment (Gr. *pros odein* 'with the song'). This suggests that the segmental structure defines the verbal contents of the message (the words), while prosody provides the music, i.e. the melody and the rhythm. Prosody comprises all properties of speech that cannot be understood directly from the linear sequence of segments. The linguistic functions of prosody are: (1) to mark off domains in time (e.g. paragraphs, sentences, phrases), (2) to qualify the information presented in a domain (e.g. as statement/terminal boundary, question/non-terminal boundary), and (3) to highlight certain constituents within these domains (accentuation) (e.g. Nooteboom 1997, Van Heuven 1994).

The expression of emotion and/or attitude is classified as yet another function of prosody. Signalling the emotional state of the speaker (e.g. happiness, sadness, anger, fear, disgust) and/or the attitude of the speaker – either towards an addressee (e.g. dominance, submissiveness) or towards the verbal contents of the message (e.g. sincerity, irony, sarcasm) are, in fact, paralinguistic (rather than linguistic) functions of prosody. They are prosodic since the signalling of emotion or attitude does not affect just a single vowel or consonant but is a property of a larger stretch of speech, spanning at least the size of an intonation domain.

The paralinguistic functions of prosody are typically subsumed under the term 'affect'. More recently, attitudinal prosody is often grouped together with emotional prosody under the superordinate term 'affective prosody' (Ross 2000), but most prior affective prosody research has focussed on emotional prosody (Fichten et al. 1992). One reason for this grouping is that attitudes and emotions are expressed by partially overlapping prosodic elements (Pell 2006).[2] However, the terms attitudinal prosody and emotional prosody are sometimes used interchangeably (Blanc & Dominey 2003, Schmitt et al. 1997, Tompkins & Mateer 1985), and it has even been commented that there is no compelling theoretical base for a distinction between attitudes such as indignation and emotions such as fear (Mozziconacci 2001). Therefore, I only use the term 'emotional prosody' to refer to the vocally expressed emotions and attitudes in order to avoid terminological inconsistency in this dissertation.

[2] According to Scherer, emotions are usually expressed in an intense way in response to a highly significant event, and the identification of emotions is largely universal. In contrast, attitudes are more enduring and concern affectively charged beliefs and predispositions. They are less intense and more socially and culturally controlled than emotions (Scherer 2003, Scherer et al. 2001).

#### 1.2.2.3 A functional view on prosody and tone

Let us define the prosodic space of a spoken language as a multi-dimensional continuum that comprises at least four (complex) dimensions, i.e. the pitch dimension (low versus high pitch, rising versus falling pitch), the loudness dimension (soft versus loud sounds, crescendo versus decrescendo), the tempo dimension (slow versus fast rate of delivery, acceleration, deceleration) and articulatory precision (clear versus sloppy articulation). There is a functional view which claims that, presumably, the prosodic space which languages may use, is finite. Therefore, if a language uses duration to mark a two-member segmental contrast between long and short vowels, the duration parameter will not play a role (or a less important role) in the marking of stress – which in other languages depends rather heavily on duration cues (Berinstein 1979, Potisuk et al. 1997, Remijsen 2002a, b). By the same token, if a language, such as Mandarin, uses pitch for lexical purposes (i.e. lexical tone), less room will be left for the signaling through pitch of paralinguistic contrasts, such as the expression of emotion. This would be a strictly functional hypothesis. If a language sacrifices one dimension of its prosodic space for the marking of lexical contrasts, it will not be possible, or at least less feasible, to use the same dimension to carry other functions. Taking a cue from Ross et al. (1986) I would predict, accordingly, that Mandarin, which uses the pitch dimension to mark a four-member lexical tone contrast, will make only limited use of the pitch dimension to also mark emotion and attitude. As a consequence of this, native listeners of Mandarin will have limited exposure to clear exemplars of prosodically expressed affect. More generally, I would predict that native listeners of a tonal language might be less intent on (and in fact less experienced in) decoding this paralinguistic use of prosody than listeners of a non-tonal language. This functional hypothesis will be tested throughout the present study.

#### 1.2.2.4 Acoustic aspects of emotional prosody

There have been ample studies which carried out acoustic analyses of emotional prosody in the past a few decades. Banse and Scherer (1996), for instance, conducted a study in which 29 acoustic features were measured. They found that F0 and mean amplitude (intensity) clearly showed the strongest connections to the emotions being produced. Other acoustic factors that are involved in production of emotional prosody are: (a) the distribution of the energy over the frequency spectrum (particularly the relative energy in the high vs. the low-frequency region, affecting the perception of voice quality or timbre); (b) the location of the formants (F1, F2…F*n*, related to the perception of articulation); and (c) a variety of temporal phenomena, including tempo and pausing (Scherer 1996).

The acoustics of emotional speech are influenced by a variety of factors. Apart from arousal and valence effects, there are other contributing factors such as talker sex, individual talker identity and emotional traits (Bachorowski & Orwen 2008: 200). Therefore, it is important that I carry out the acoustic analysis of the chosen emotional prosodies in a more integrated way. The specific acoustic correlates which speakers and listeners use in the production and the perception of vocal emotions will be described in Chapter 6.

### 1.2.3 The six chosen emotional prosodies

There are six emotional prosodies chosen for in the dissertation: 'neutrality', 'happiness', 'anger', 'surprise', 'sadness' and 'sarcasm'. 'Neutrality' is considered as no emotion, for example: news-style. 'Happiness' and 'anger' in the present study both refer to hot happiness and hot anger. The reasons to choose these six emotions are:

(1) 'Neutrality' is chosen for being a point for comparison, such that other emotions need to differ from 'neutrality' to be considered as an emotion. It will also help to draw an acoustic picture of other emotions at the later stage of the study.
(2) 'Happiness', 'anger' and 'sadness' are traditionally studied in previous studies, as they are arguably the basic emotions of human communication (Darwin 1872).
(3) Strictly speaking, 'surprise' and 'sarcasm' are not emotions, but attitudes. However, in Mitchell and Ross's (2013) review, 'surprise' is sometimes considered to be a function of emotional rather than attitudinal prosody (Monrad-Kohn 1947, 1963). 'Surprise' has been studied before and it has been claimed by some researchers (e.g. Yip 2006) that many tonal languages use rising intonation to express surprise. Therefore, I am interested in finding out whether Chinese also uses rising intonation to portray 'surprise' as is implied by Yip.
(4) Through observation, 'sarcasm' is often used to express annoyance, cold anger or complicated negative feelings in Chinese culture. It is used frequently in Chinese everyday communication. However, it has not been properly studied previously. Therefore, I chose this emotional prosody in this dissertation to find out more about it.

## 1.3 Research questions

Specifically, in this dissertation I will aim to find answers to the following questions:

(i) How well can native Chinese, Dutch naïve listeners and advanced Dutch learners of Chinese perceive the six Chinese emotional prosodies vocally portrayed by Chinese native speakers? What will be the confusion patterns of the three listener groups?
(ii) How well can native Chinese, Dutch naïve listeners and advanced Dutch learners of Chinese perceive the six Chinese emotional prosodies vocally portrayed by Dutch L2 speakers of Chinese? What will be the confusion patterns of the three listener groups?
(iii) Can Dutch L2 speakers of Chinese produce emotional prosodies in their L2 as well as they do in their L1 – Dutch? What will be the similarities and differences between these two types of production?
(iv) Does L2 limit the expression of emotional prosody, especially when the native language of the L2 speakers of the tonal language is a non-tonal language?
(v) Is the functional view true, predicting that listeners of a tonal language might be less intent than listeners of a non-tonal language on (and in fact less experienced in) decoding the paralinguistic use of prosody?
(vi) What acoustic parameters contribute to differentiate between emotional prosodies in general? What acoustic correlates do speakers and listeners use to produce and

perceive the vocal emotions in their L1 and in an L2? Do Dutch L2 speakers of Chinese use L1-transfer when producing emotional prosody in Chinese? To what extent does automatic recognition reflect the perception of the emotional prosodies by the human listeners?

(vii) Is the in-group advantage universal, claiming that listeners are better in recognizing emotional prosody produced in their native language than in their L2 or an unknown language? Moreover, is the perception of vocal emotion cross-culturally symmetrical between Chinese and Dutch listeners, i.e., will Dutch and Mandarin listeners have similar abilities of identifying emotional prosody expressed in the other language?

(viii) Are perception and production of emotional prosody universal? Or are they rather more language-specific and culture-specific?

## 1.4 Research approach

In order to answer the research questions, I will run three judgment studies (more detailed information about the experimental design and procedures will be provided in Chapter 2). The first judgment study includes one perception experiment (Exp 1), in which native Chinese listeners, naïve Dutch listeners and advanced Dutch learners of Chinese perceive and identify the six Chinese emotional prosodies portrayed by native Chinese speakers. This experiment aims to find an answer to research question (i): how well do the three listener groups perceive the native-Mandarin produced emotional prosodies. The results will be used as the base-line condition for later studies. The second judgment study includes two perception experiments: the first perception experiment where the same listener groups listen to the same six Chinese emotional prosodies but produced by Dutch L2 speakers of Chinese (Exp 2A), is designed to find answers to research question (ii), i.e. a) how well can the three listener groups perceive the six Chinese emotional prosodies vocally portrayed by Dutch L2 speakers of Chinese? b) what are the confusion patterns of the three listener groups? In the second perception experiment (Exp 2B), Dutch native listeners will listen to the same six emotional prosodies portrayed in their native language (Dutch) by the same Dutch L2 speakers of Chinese. This is to test how well the same Dutch L2 speakers of Chinese produce the emotional prosodies in their L1.[3] The results of this perception experiment will be compared with the results obtained in the first perception experiment of the second judgment study to answer research questions (iii) and (iv). Research question (v) will be answered after the first and the second judgment study, questioning whether the functional view is true. There will be an acoustic analysis based on selected stimuli after I run the two judgment studies. The results will answer the research question (vi). The third judgment study will be conducted in a reciprocal way. It includes two perception experiments in which Chinese and Dutch novice listeners perceive the six emotions vocally portrayed in their L1 and in the other language (Exp 3). This experiment is

[3] According to Flege's Speech Learning Model, speaking Mandarin as a foreign language may have compressed the speakers' realisation of emotions in their L1. It might be the case in the present study. However, I am interested in the difference between two types of production of the vocal emotions (one is in speakers' L2; the other is in their L1). Therefore, Flege's model will not influence the results, as the results are going to be relative.

designed to test whether the in-group advantage claimed by other researchers is universal, which would answer research question (vii). The three judgment studies altogether will answer the research question (viii): are perception and production of emotional prosody universal? Or are they rather more language specific and/or culture specific?

### 1.5 Thesis outline

This dissertation comprises the description of a series of perception experiments investigating the research questions outlined above. Chapters 3 to 7 have their own introduction and conclusion sections, since they have been written as independent articles. Therefore, there are unavoidable overlaps between the introductory sections of these chapters, as well as with the general introduction, the background and the explanation of the experimental design and procedures. Chapter 2 will provide a literature review of what other researchers have contributed to answering the above-mentioned research questions and detailed information of how I planned, designed and conducted the three judgment studies. In Chapter 3, I will report the results of the first judgment study, which was designed to examine how well native Chinese listeners, naïve Dutch listeners and advanced Dutch learners of Chinese perceive and identify the six Chinese emotional prosodies portrayed by native Chinese speakers. I will also present confusion matrixes of the three listener groups and other results. In Chapter 4, I will show the results of the second judgment study, in which the same listener groups perceive the six Chinese emotional prosodies but produced by Dutch L2 speakers of Chinese. Chapters 3 and 4 will answer the research question (ii). Chapters 3 and 4 have been accepted as articles by two peer-viewed journals; therefore, these two papers will be included in the dissertation independently. In Chapter 5, I will present the full data of the first and the second judgment studies. This chapter has been written in a comparative manner from the production point of view. There will also be a speaker-listener combination study in the same chapter. Chapter 5 will show a complete picture of the differences between the productions in L2 speakers' L2 and in their L1. Research questions (iii) and (iv) will then be answered. I will give the answer to research question (v) – whether the functional view is true, after the first and the second judgment study. An acoustic analysis and automatic recognition of the various emotional prosodies will be carried out to answer the question (vi) in Chapter 6. The acoustic analysis will contain selected stimuli of the six Chinese emotional prosodies produced by L1 and L2 speakers, as well as Dutch emotional prosodies expressed by the same L2 speakers of Chinese. There will be a degree of overlap between Chapter 5 and the previous two chapters (Chapters 3 and 4). Chapters 5 and 6 together will form a long article which will be submitted to a journal as a single article. Chapter 7 will report the test of the in-group advantage. The results will answer research question (vii). This chapter will be written as an independent article and later submitted to a journal. The final chapter, Chapter 8, will summarize what I found in the three judgment studies, including some unexpected findings. The three judgment studies together will answer research question (viii). Moreover, I will provide the possible explanations of unexpected findings and make suggestions for future research.

# Chapter Two

# Background and Methodology

## 2.1 Introduction

In this chapter I will go through a literature review of what previous studies have done in terms of perception and production of emotional prosody by native and non-native listeners/speakers. Apart from that, I will also provide detailed information on research methods, including how I conducted the present study to answer my research questions and how I carried out the perception experiments in the present study, as well as how I analyzed the production of the six emotional prosodies portrayed in my speakers' L1 and L2.

## 2.2 Background

### 2.2.1 General introduction

Darwin (1872) was the first to claim that affective expressions, including those produced via the vocal channel, are veridical. He generally showed possible veridical associations between vocal acoustics and the vocalizer's emotional state. He also pointed out two aspects which very much influenced later studies in the field: (1) vocal signals can trigger emotional responses in listeners; (2) these signals can elicit learned emotional responses. These two observations are fundamental to any research on perception of vocally portrayed emotion, as they formally claim that there is a link between acoustic signals and emotions so that researchers since then were able to deepen the knowledge of how vocalizer (speaker) and recipient (listener) produce and perceive emotion via the audio-channel (Bachorowski & Owren 2008). Moreover, many previous studies have confirmed the two observations by Darwin, claiming that information concerning emotional state is encoded in vocal acoustics and subsequently decoded by listeners in order to respond to the speaker's emotional state (e.g. Juslin & Laukka 2001, Scherer 1986, 2003). They also found that specific patterns or configurations of vocal cues are reliably associated with different emotional/affective states (e.g. Borden & Harris 1994, Deller et al. 1993, Scherer 1986, Spackman et al. 2009). Furthermore, Bachorowski and Owren's (2008: 196) review states that there is now a substantial body of work focused on how emotion is conveyed by and perceived from vocal acoustics. Although this research has arguably not enjoyed the same degree of cumulative success as has work on the communication of emotion through the facial channel, there is nonetheless a solid body of evidence showing that specific vocal acoustic features are reliably associated with affect-related arousal (or activation) on the part of vocalizers, and that listeners in turn can reliably perceive arousal from vocal

acoustics. These claims together show that the audio channel plays an important role in the perception and production of emotion.

It is good to know some general background of perception of vocally produced emotion in listeners' native language (L1) before reviewing the perception of emotional prosody produced in listeners' non-native languages (L2). According to previous studies, humans can accurately decode discrete emotions from speech-embedded prosody at levels well above chance (e.g. Banse & Scherer 1996, Biehl et al. 1997, Haidt & Keltner 1999, Juslin & Laukka 2001, 2003, Rosenberg & Ekman 1995). Moreover, Johnstone and Scherer's (2000) review mentions that when listeners are asked to identify the intended emotion in utterances produced by actors, accuracy is significantly better than chance – although at a moderate level overall, typically about 55%. According to Bachorowski and Owren's (2008) review, identification rates are usually best for anger, fear, and sadness. Results for positive emotions have varied, but in an informative way. Accuracy is typically high when listeners are given only one positive response option (e.g., Johnson et al. 1986, Scherer et al. 1991). However, correct responses drop significantly when other positively toned options are tested, such as 'elation', 'contentment' or 'interest'. A similar effect may contribute to the identification of 'sadness', which is sometimes the only low-arousal option offered among the negative emotions. Another factor which might affect listener recognition accuracy is the experimental design, such as whether one should use forced-choice procedures or free-choice tests. It is concluded by previous researchers (e.g., Johnson et al. 1986, Pakosz 1983) that forced-choice procedures generate better recognition results than free-choice ones. In real human communication, there are other possible factors that can influence listener identification accuracy, for example, the speaker's age, sex and style, the listener's personal interpretation of the emotion label, the testing method, etc.

### 2.2.2 Perception of emotional prosody in L1, L2 or an unknown language

There are a few studies which touch on the perception of emotional prosody by both native and non-native listeners. To some extent, previous findings all claimed that perception of emotion by different culture groups is partly universal and partly language/culture-specific. For instance, Van Bezooijen (1984) studied ten emotional prosodies: neutral, happy, sad, anxious, angry, afraid, surprised, disgusted, annoyed and embarrassed. Her study aimed to find out how (Taiwanese) Chinese and Japanese listeners, who did not have any knowledge of Dutch, perceived Dutch emotional prosodies at the sentence level. Perceptual experiments showed that Dutch native listeners got the highest correct identification rate and Japanese listeners performed poorest. But both of the Asian listener groups performed well above chance level. Studies by Tickle (2000) and by Scherer et al. (2001) both further claim that listeners can recognize emotional prosody in an unknown language better than chance level but the misidentification rates increase as speaker and listener languages become more dissimilar. Graham et al. (2001) examined the ability of native and non-native speakers of English to identify emotions being portrayed by English speakers. They concluded that the ability to accurately identify emotions being portrayed through vocal cues in a second language (L2) may not be acquired by L2 learners without extensive exposure in a native context or without special attention to developing these skills in an

instructional context. Moreover, an analysis of judgments made by learners of English as a Second Language (ESL) at different proficiency levels did not show an increase in ability to judge the emotional content of English speech with increased language proficiency. Thompson and Balkwill (2006) conducted a similar experiment in which twenty English-speaking listeners judged the emotive intent of utterances spoken by male and female speakers of English, German, Chinese, Japanese, and Tagalog. Identification accuracy was above chance for all emotions expressed in all languages. Across languages, sadness and anger were more accurately recognized than joy and fear. The (English) listeners showed an in-group advantage for decoding emotional prosody, with highest recognition rates for English utterances and lowest rates for Japanese and Chinese utterances. This would indicate that, again, emotional prosody is decoded by a combination of universal and culture-specific cues. Shoshi and Gagnié (2010) investigated differences in the perception of six culturally encoded French social affects through audio and visual channels for French native listeners, naïve Japanese listeners and trained Japanese learners of French. The trained Japanese learners of French recognized the emotions better than the naïve Japanese listeners; however, culture-specific attitudes (i.e. suspicious irony and obviousness) were confused by Japanese listeners (including trained listeners). Facial information cues seem to be more salient here than auditory cues.

### 2.2.3 Production of emotional prosody in speakers' L1 and L2

In Bachorowski and Owren's (2008) review, from the production point of view, Darwin's two observations (1872) form the core of an 'affect induction' view of vocal signaling, which began as a functional account of nonhuman primate calling (Owren & Rendall 1997, 2001, Rendall & Owren 2002), but may also apply to affect-related vocal signaling in humans (Owren et al. 2003). In sum, the affect-induction perspective argues that vocal expressions of emotion are not displays of vocalizer states as much as they are tools of social influence. This view shows that a listener's personal interpretation of the incoming signals from the speaker determines how successful the communication will be. Mismatches between a speaker's intended emotion and a listener's perceived emotion can cause a failure in the paralinguistic communication. The success of a communicative encounter requires not only the ability to convey one's own affect but also the ability to accurately gauge that of the other person. This process does not always work well, particularly when engaging in telephone communication, where face-to-face contact is not possible and the reading of affect is solely dependent on the auditory channel of communication (Mitchell & Ross 2013). Therefore, investigating how speakers encode emotional prosody will guide us to find out what vocal cues speakers use to vocally express emotion (or to trigger a listener's own affective response) and why these cues sometimes can be misinterpreted or mismatched to another affective state by listener.

Since Darwin (1872) claimed that there is a direct correspondence between particular signaler states and the communicative display produced, there were researchers who further developed the view, claiming that each emotion is associated with distinctive acoustic cues (e.g. Banse & Scherer 1996, Leinonen et al. 1997, Scherer 1986, 1989). Specifically, Banse and Scherer (1996) analyzed the vocalizations generated by 12

professional actors who each portrayed 14 emotions. In this study, 29 acoustic features were measured. They found that fundamental frequency (F0) and mean amplitude clearly showed the strongest connections to the emotions being portrayed. According to Scherer's (1996) review, there were other acoustic cues which also contributed to the production of emotional prosody: (a) the vocal energy (or intensity, perceived as loudness of the voice); (b) the distribution of the energy in the frequency spectrum (particularly the relative energy in the high vs. the low-frequency region, affecting the perception of voice quality or timbre); (c) the location of the formants (F1, F2…F*n*, related to the perception of articulation); and (d) a variety of temporal phenomena, including tempo and pausing; (e) F0 variability (including both overall range and moment-to moment perturbations, e.g. jitter). However, recently researchers not only studied jitter (or related measures) but also HNR (Harmonics to Noise Ratio) to better understand some particular emotions, for example, sadness.

Although there were plenty of studies on production of emotional prosody, previous studies mainly focused on the vocal production of emotion by native speakers from one particular linguistic group. There was little research on production of emotional prosody in an L2, especially when the L2 is a tonal language, e.g., Mandarin, Thai, or Vietnamese. The only related literature that can be found so far is Anolli et al.'s (2008) study, which directly studied the topic of vocal production of emotion cross-culturally. They conducted research on vocal production of emotion by Chinese and Italian young adults. They confirm that different emotions may be expressed through variations in the modulation of vocal cues, in both cultures; on the other hand, differences in the specific patterns of vocal cues in expressing emotions were identified between Chinese and Italian participants. However, there is no earlier study which directly dealt with production of emotional prosody in speakers' L1 and in their L2. Therefore, the production part in the present study is a pioneering investigation on how well native speakers of a non-tonal language (e.g., English or Dutch native speakers) can express emotional prosody in a tonal L2.

## 2.3 Methods

### 2.3.1 Overview

In order to answer the research questions, I designed three judgment studies using an experimental approach in this dissertation. The first judgment study includes one perception experiment in which native Chinese listeners, naïve Dutch listeners and advanced Dutch learners of Chinese identify six Chinese emotional prosodies portrayed by native Chinese speakers. It aims to answer the first research question: how well do native and non-native listeners, including naïve listeners and L2 learners of the target language, perceive the emotional prosodies portrayed by native speakers? In this judgment study, apart from looking at the correct recognition rates of the three listener groups, I will also present confusion matrices and confidence ratings of the three listener groups. This will help us to obtain a clearer picture of what the differences are between native and non-native listeners in perceiving emotional prosody and whether having high language proficiency in the target language will help L2 learners of the target language perceive emotional prosody portrayed better in that language. The

results will serve as the control group in the entire research. The second judgment study consists of two perception experiments: in the first, the same three listener groups perceive the same six Chinese emotional prosodies but now produced by Dutch L2 speakers of Chinese. It is designed to find out: (i-a) compared to the native Chinese speakers (the control group), how well can the Dutch L2 speakers of Chinese vocally produce the Chinese emotional prosodies (the judgment will be based on the correct identification rates of the three listener groups); (i-b) what are the differences between listeners perceiving emotional prosody portrayed by native and L2 speakers of the target language.

In the second perception experiment, Dutch native listeners who do not have any knowledge of Chinese, will identify the same six emotional prosodies portrayed in their native language (Dutch) by the same Dutch L2 speakers of Chinese. This experiment is carried out to test how well the Dutch L2 speakers of Chinese can vocally produce the same emotional prosody in their native language. Moreover, it can tell us how Dutch native listeners perceive emotional prosodies generated in their L1. The results of this perception experiment will be compared with the results obtained in the first perception experiment. There will be an acoustic analysis based on selected stimuli after I run the two judgment studies. The results will reveal the vocal correlates that L1 and L2 speakers use to produce emotional prosody in their L1 and L2.

In sum, the second judgment study and the acoustic analysis will answer the following research questions: (ii) How well do native Chinese, Dutch naïve listeners and advanced Dutch learners of Chinese perceive the six Chinese emotional prosodies vocally portrayed by L2 speakers of Chinese? (iii) Can L2 speakers of Chinese vocally produce emotional prosodies in their L2 as well as they do in their L1? What will be the similarities and differences between the two productions? (iv) Does L2 limit the expression of emotional prosody, especially when the native language of L2 speakers of the tonal language is a non-tonal language? (vi) What acoustic parameters contribute to differentiate between emotional prosodies in general? What acoustic correlates do speakers and listeners use to produce and perceive the vocal emotions in their L1 and in an L2? Do Dutch L2 speakers of Chinese use L1-transfer to produce emotional prosody in Chinese? To what extent does automatic recognition reflect the perception of the emotional prosodies by the human listeners?

After the first and the second judgment study, research question (v) will be answered, i.e., is there any support for the functional view, predicting that listeners of a tonal language might be less intent on (and in fact less experienced in) decoding the paralinguistic use of prosody than listeners of a non-tonal language?

The third judgment study includes one perception experiment in which Chinese and Dutch novice listeners of each other's language perceive the six emotional prosodies portrayed in Chinese and Dutch by native speakers of the respective languages. This experiment will be conducted in a reciprocal way and it is designed to test whether the in-group advantage claimed by other researchers is universal, and to investigate whether or not the ability of Chinese and Dutch native listeners to perceive emotional prosody in the other (unknown) language is symmetrical. The third judgment study will answer research question (vii).

### 2.3.2 Speakers

#### 2.3.2.1 Acted stimuli vs. natural stimuli

According to previous studies, it is suggested that using acted stimuli is better than natural stimuli (Scherer 2003). According to the review of Bachorowski and Owren (2008), the obvious problem with relying on acted samples is that these may not necessarily correspond to naturally produced vocalizations. One counterargument is that much of our verbal communication involves making impressions on others, and so having vocalizers act 'as if' they were experiencing a particular state is not markedly different from natural communicative circumstances. However, when the issue taken together with evidence from natural emotion-inducing circumstances showing that individual variability in vocalizer acoustics can be quite substantial (e.g., Streeter et al. 1983), it may be the case that the careful analysis of acoustic cues to acted emotion provides more information about emblematic portrayals of affective states than about naturally occurring cueing (Scherer 2003).

In addition, it has been seen quite often in previous studies that some individual actors (especially females) were much more convincing in their portrayals than others (e.g., Leinonen et al. 1997, Pell 2001, Scherer et al. 1991, Schröder 2000, Sobin & Alpert 1999, Walbott & Scherer 1986). Therefore, I decided to use four native speakers who are amateur actors/actresses and four non-native speakers to vocally produce the six intended emotional prosodies in their L1/L2 for all the perception studies, as it would help to balance off this kind of sex and talker differences in the production of stimuli. And I will use carefully selected and acted samples produced by the (non-)native speakers for acoustic analysis.

#### 2.3.2.2 Native speakers of Chinese

In the present study, there were four native Mandarin Chinese speakers (2 males and 2 females, mean age = 45), who voluntarily produced the six emotional prosodies in spoken Chinese for the perception experiments. The four native Chinese speakers were amateur actors/actress who had acting training and stage performance experience.

#### 2.3.2.3 Dutch L2 speakers of Chinese

Four Dutch L2 speakers of Chinese (2 males, 2 females, mean age = 33 years) voluntarily participated in the recording of the stimuli for the second judgment study. These four Dutch L2 speakers of Chinese were teachers from the Chinese department of Leiden University in the Netherlands. None of them were early bilinguals. They had learnt Chinese for 6 to 10 years; they had been teaching Chinese for 2 to 10 years when the recordings were made. All spent at least one year living or studying in mainland China or Taiwan.

### 2.3.3 Listeners

In the first and second judgment studies, 20 native Mandarin listeners (10 males, 10 females, mean age = 24 years), 20 naïve Dutch listeners (10 males, 10 females, mean age = 33 years) and 20 advanced Dutch learners of Chinese (10 males, 10 females, mean age = 20 years) voluntarily participated in each of the perception experiments, where they were required to perceive the emotional prosodies portrayed by L1 and L2 speakers of Chinese, respectively. The Chinese listeners were bachelor and master students at the University of Science and Technology Beijing who hailed from different parts of China. All spoke Mandarin Chinese as their mother tongue. The naïve Dutch listeners were mainly bachelor students at the Humanities Faculty of Leiden University in the Netherlands and volunteers with variable education backgrounds. None of the naïve Dutch listeners spoke any Mandarin. The advanced Dutch learners of Chinese were mainly third-year BA students in the Chinese Program of Leiden University; the others were MA students and some outstanding second-year BA students. Early bilinguals were excluded; therefore, all students had learnt Mandarin after the age of eighteen. There was no special course in the curriculum designed for training these students to recognize emotions in Chinese.

In the third judgment experiment, 20 native Chinese listeners (10 males and 10 females) and 20 native Dutch listeners (10 males and 10 females) who did not know each other's language (novice listeners) voluntarily participated in the perception experiment in which they were required to identify the six emotional prosodies ('neutrality', 'happiness', 'anger', 'surprise', 'sadness' and 'sarcasm') portrayed in Dutch by native Dutch speakers, who were the same individuals as the Dutch L2 speakers of Chinese in the second judgment study.[4] This experiment was designed to test how well novice Chinese and Dutch native listeners perceive emotional prosody portrayed in the other language. The results would tell us whether the in-group advantage claimed by other researchers applies to these two cultural groups and whether the Chinese and Dutch native listeners possess similar ability of identifying emotional prosody correctly in an unknown language. For the aim of the third judgment study, the advanced Dutch L2 leaners of Chinese were excluded, even though they were also native Dutch listeners.

[4] These 20 native Chinese and Dutch (novice) listeners were drawn from the same population group as the 20 Chinese native listeners and the 20 naïve Dutch listeners in the first judgment study. The third judgment study was conducted three months later after the first two judgments were carried out.

### 2.3.4 Materials and procedure

#### 2.3.4.1 The first judgment study

The first judgment study only included one perception experiment, which was set up to test how well the Chinese control group vocally expressed the six emotions in Chinese. In the first judgment study, I used six Mandarin statements (e.g. *She is three months pregnant; He has been to Xiao Ge's place once*). The requirements for the stimulus selection were: (1) stimuli contain all the tones in Mandarin, i.e. 'high-level tone', 'rising tone', 'falling-rising tone', 'falling tone' and 'neutral tone' (e.g., Howie 1976); (2) stimuli have to be semantically neutral but can easily be expressed with different emotions; (3) both short and longer sentences have to be included, in case utterance length might play a role in the perception of emotional prosody. The list of stimulus sentences for the first judgment study is shown in Table 2.1. The results of the first judgment study will be used as the control group.

Table 2.1. *Stimulus list in Chinese (Pinyin) with English glosses.*

| | |
|---|---|
| 1. | **Shì nǐ.*<br>'It is you.' |
| 2. | *Xièxiè nǐ.*<br>'Thank you.' |
| 3. | *Xiǎo wáng wánquán bù zhīdào zhè jiàn shì.*<br>'Xiao Wang does not know about this matter.' |
| 4. | *Jīntiān xiàwǔ tā bùnéng lái cānjiā zhège huì.*<br>'He cannot attend the meeting this afternoon.' |
| 5. | *Tā huáiyùn sān ge yuèle.*<br>'She is three months pregnant. ' |
| 6. | **Tā qùguò xiǎo gě jiā yì cì.*<br>'He has been to Xiao Ge's place once.' |

Note: '*' means sentences were excluded in the second perception experiment of the first judgment study. Macron '¯' = high-level tone, acute accent '´'= rising tone, haček 'ˇ' = falling-rising tone, grave accent '`' = falling tone; a syllable without tone mark has neutral tone.

In the perception experiment, each of the six Mandarin statements was vocally expressed in six different emotions (neutrality, happiness, anger, surprise, sadness and sarcasm) by the four native Chinese speakers. The stimuli were digitally recorded (44.1 KHz, 16 bits) in a sound-proofed booth through a Logitech desk-top microphone. This procedure resulted in a stimulus set that consisted of 6 Chinese statements × 4 Mandarin speakers × 6 emotions = 144 discrete emotional utterances.

It is concluded by previous researchers that forced-choice procedures generate better recognition results than free-choice tasks (e.g. Johnson et al. 1986, Pakosz 1983). Therefore, all the participants (including native Chinese listeners, naïve Dutch listeners and advanced learners of Chinese) were asked to make a forced choice of the speaker's intended emotion, from the six given emotions, immediately after they heard a stimulus. They also gave a confidence rating to each choice they made. A three-level confidence

rating scale was used, with the following interpretation: 3 = 'The speaker expressed the intended emotion well. I am very confident in my answer', 2 = 'The speaker reasonably expressed the intended emotion. But I am not so sure about my answer' and 1 = 'The speaker did not express the intended emotion well. I made the choice only by guessing.' The confidence scale was introduced in order to obtain a potential weighting factor such that responses given with great confidence would be weighted more heavily than responses that were largely based on guessing.

#### 2.3.4.2 The second judgment study

The second judgment study included two perception experiments: the first perception experiment was carried out to test how well the three listener groups, i.e. native Chinese, naïve Dutch listeners and advanced Dutch learners of Chinese, perceived Chinese emotional prosody expressed by the Dutch L2 speakers of Chinese. It would thus reveal how well the Dutch L2 speakers of Chinese encoded the six emotions in Chinese, compared to the control group. The second experiment was conducted to determine how well Dutch L2 speakers of Chinese produced the same emotional prosodies in their mother tongue – Dutch.

In the first perception experiment, the four Dutch L2 speakers of Chinese were asked to express the six emotional prosodies in Chinese. The stimuli were digitally recorded under the same conditions as in the first judgment study. Two sentences were discarded from the stimulus set (see Table 2.1), as these two sentences were less well perceived by the three listener groups in the first perception test. Therefore, the final stimulus set for the second perception experiment consisted of 4 Chinese statements × 4 Dutch L2 speakers × 6 emotions = 96 discrete emotional utterances. It made this experiment shorter than the perception experiment in the first judgment study. I only processed and showed the shared data in later comparisons. The three listener groups, including native Chinese, Dutch naïve listeners and advanced Dutch learners of Chinese, repeated the same experiment procedure of the first judgment study in this test round.

In the second perception experiment, 20 native Dutch listeners perceived the six emotions produced by the same four Dutch L2 speakers of Chinese, but in their mother tongue – Dutch.[5] This experiment was designed to find out how well the Dutch L2 speakers of Chinese produced the same emotional prosodies in their L1. The four Mandarin statements used in the first perception experiment were then translated into Dutch by the four Dutch L2 speakers of Chinese where sentence length, syntactic structure, syllables and sentence meaning were well controlled. Therefore, the final stimulus set for the second perception experiment consisted of 4 Dutch statements × 4 Dutch L2 speakers of Chinese × 6 emotions = 96 discrete emotional utterances. The list of stimulus sentences for the second judgment study is shown in Table 2.2. The same procedure as in the first judgment study was used to obtain the judgments. In fact, some of the sentences may be associated more readily with some emotions than with

[5] These 20 native Dutch listeners were drawn from the same population as the 20 naïve Dutch listeners in the first judgment study. This second perception experiment was conducted three months after the first perception experiment was carried out.

others but on aggregate the lexico-syntactic materials will not be biased towards specific emotions.

Table 2.2. *Stimulus list in Dutch with Broad IPA transcriptions and English glosses.*

| | |
|---|---|
| 1. | *Dank je wel.*<br>dɑŋk jə ʋɛl<br>'Thank you.' |
| 2. | *Xiaowang weet dat helemaal niet.*<br>ʃau ʋɑŋ ʋetɑt heləmal nit<br>'Xiao Wang does not know about this matter.' |
| 3. | *Vanmiddag kan hij niet naar de vergadering.*<br>ʋɑmɪdɑχ kɑni nit nar də ʋərɣadərɪŋ<br>'He cannot attend the meeting this afternoon.' |
| 4. | *Zij is drie maanden zwanger.*<br>zɛɪ ɪs dri mandə zʋɑŋər<br>'She is three-months pregnant.' |

#### 2.3.4.3 The third judgment study

The third judgment study was designed to test whether the in-group advantage found by other researchers (e.g. Kilbride & Yarczower 1983, Markham & Wang 1996) is universal, claiming that native listeners are generally better at identifying emotional prosody in their L1 than in an unknown language. The third judgment study also aims to find out whether the ability to identify emotional prosody in an unknown language cross-culturally is symmetrical.

The third judgment study was conducted in a reciprocal way. In the reciprocal approach both culture groups A and B perceive emotional prosody not only expressed in their own native language (A > A, B > B) but also emotions expressed in the other language (A > B, B > A). As an example of the latter situation, English listeners may recognize emotional prosody in Japanese, and vice versa. Although some studies (e.g. Albas et al. 1976, Dennis 1982, Gitter et al. 1972) used this reciprocal approach, the two cultural groups involved were ethnically different rather than culturally-or-linguistically dis-similar. Other studies also adopted this method (e.g. Ekman 1972); however, they only investigated the perception of facially expressed emotions between two culture groups. Even though previous studies have clearly indicated an in-group advantage in the perception of emotion cross-culturally, the reciprocal method was not used so that those studies present an incomplete picture, especially when it comes to the perception of vocal emotion. Therefore, I would like to conduct the third judgment test applying the reciprocal method to the two cultural groups, i.e. Chinese and Dutch.

In this study, 20 Chinese and 20 Dutch native listeners who did not know each other's language (novice listeners) perceived the six emotional prosodies ('neutrality',

'happiness', 'anger', 'surprise', 'sadness' and 'sarcasm') portrayed in their L1 and in the other language by Chinese and Dutch native speakers.[6] The same experimental materials and procedures of the first two judgment studies were adopted in the reciprocal manner in this study (details in Chapter 7). The results and conclusions are also presented in Chapter 7.

### 2.3.5 Acoustic analysis

An acoustic analysis was conducted after the first and the second judgment studies. Two sentences were discarded from the stimulus list (see Table 2.1), as they were not well perceived by the listener groups in the first judgment study (the control group). The optimized stimulus list resulted in three sets of stimuli for the acoustic analysis in which each set consisted of 4 (non-)native speakers × 4 sentences × 6 emotional prosodies = 96 stimuli. Therefore, there were 288 stimuli in total (96 × 3 = 288) for the acoustic analysis: emotional prosody portrayed in L1 Chinese, emotional prosody portrayed in L2 Chinese by Dutch L2 speakers of Chinese, emotional prosody portrayed in L1 Dutch. First I extracted pitch contours of all the utterances, and then extracted selected other variables (see below). Finally, I compared the individual acoustic parameters between sets. In this manner, one can see the influence of each parameter in the vocal production of emotion. It also allows us to see what vocal correlates L1 and L2 speakers use in portraying emotional prosody in their L1/L2. Moreover, it will show us which parameters are relatively more important than others in the production and perception of emotional prosody, and which parameters can adequately differentiate between emotions.

In order to analyze the acoustic parameters which might contribute to the production of the emotional prosodies, I took Scherer's (1996) review as a point of reference. In addition, I investigated some parameters which were not mentioned in his review. Therefore, I studied the following acoustic variables obtained from the computer analyses of the speech signals: (a) tempo (duration/time); (b) mean fundamental frequency (F0) and F0 in the first and last quarters of the utterance duration, which is named 'F0 slope' in this dissertation, as well as standard deviation of F0; (c) distribution of the energy in four contiguous frequency bands (d) vocal energy (mean intensity standard deviation); (e) mean jitter; (f) mean HNR (Harmonics to Noise Ratio).

Automatic recognition was attempted on the basis of the acoustic analysis. The eight above-mentioned variables were entered into a Linear Discriminant Analysis (LDA) to identify the six emotional prosodies portrayed by the three speaker groups, i.e. L1 Dutch speaker, L2 Mandarin speakers and L1 Mandarin speakers (where the former two groups are the same individuals). This automatic recognition may reflect the variables used by the human listeners. The acoustic analysis and the automatic recognition will be presented in Chapter 6.

[6] Some data obtained in the first and second judgment study were re-used in the third judgment study.

# Chapter Three

# Perception of Chinese Emotional Prosody by Chinese Natives, Naïve Dutch Listeners and Dutch L2 Learners of Chinese

**Abstract**

This chapter investigated the perception of six Chinese emotional prosodies (neutrality, happiness, anger, surprise, sadness and sarcasm) by 20 Chinese native listeners, 20 naïve Dutch listeners and 20 advanced Dutch L2 learners of Chinese.[7] The results showed that advanced Dutch L2 learners of Chinese recognized Chinese emotional prosody significantly better than Chinese native listeners and Dutch naïve listeners. The results also indicated that naïve non-native listeners could recognize emotions in an unknown language as well as the natives did. Chinese native listeners did not show an in-group advantage for identifying emotions in Chinese more accurately and confidently. 'Neutrality' was the easiest emotion for all the three listener groups to identify and 'anger' was recognized equally well by all the listener groups. The prediction made in the beginning of the study is confirmed, which claims that listeners of a tonal language will be less intent on paralinguistic use of prosody than listeners of a non-tonal language. The results in this chapter will be used as the control group for Chapter 4.

[7] This chapter has appeared as Y. Zhu (2013a). Which is the best listener group? Perception of Chinese emotional prosody by Chinese natives, naïve Dutch listeners and Dutch L2 learners of Chinese, *Dutch Journal of Applied Linguistics*, 2, 170–183.

### 3.1 Introduction

It is customary to distinguish between segmental and suprasegmental (also called prosodic) aspects of speech. Segmental properties are inherent properties of individual vowels and consonants or co-intrinsic properties that are predictable from the sequence of segments. Prosody then refers to the ensemble of properties of speech that cannot be derived from the mere sequence of segments (e.g. Lehiste 1970, Nooteboom 1997). It follows from this definition that prosody includes the temporal and melodic effects of lexical tone, word-stress, phrasing, accentuation and intonation, as well as the articulatory setting and voice quality of longer stretches of speech. Prosody fulfils both linguistic and paralinguistic functions in speech communication. Linguistic functions of prosody would be to divide the stream of speech into chunks of information that should be processed as meaningful units, highlight specific words as communicatively important, or differentiate between sentence types such as question and statement (Grandjean et al. 2006). A paralinguistic function of prosody is to signal the emotional state of the speaker (e.g. happiness, sadness, anger, fear, disgust) and/or the attitude of the speaker – either towards an addressee (e.g. dominance, submissiveness) or towards the verbal contents of the message (e.g. sincerity, irony, sarcasm). These paralinguistic functions of prosody are often subsumed under the term 'affect'. However, I only use the term 'emotional prosody' in this article to refer to the vocally expressed emotions and other affects. Moreover, although it has been shown that changes in articulatory setting, such as lip spreading (smiling) by a happy speaker (Tartter & Braun 1994), and in voice quality, such as a rough voice during anger (Grandjean et al. 2006), also contribute to the expression of emotion, I will concentrate on the use of speech timing and melody as the primary correlates of emotion.

A lot of attention has been drawn to vocal expressions of emotion and their acoustic accounts since Darwin concluded that affective expressions, including those produced via the vocal channel, are veridical (Darwin 1872). However, most of the previous studies within psycholinguistic and phonetics focused on human perception and use of different emotional prosodies of one particular language (e.g. Albas et al. 1976, Ladd et al. 1985). There were only few researchers who studied the perception of emotional prosody by both native and non-native listeners. For instance, Van Bezooijen (1984) studied ten emotional prosodies: neutral, disgust, surprise, shame, interest, joy, fear, contempt, sad, and angry. Basically, her study aimed to find out how (Taiwanese) Chinese and Japanese listeners, who did not have any knowledge of Dutch, perceived Dutch emotional prosodies at the sentence level. Perceptual experiments showed that the three listener groups were able to recognize the emotional prosodies well above chance level. Dutch native listeners got the highest correct identification rate and Japanese listeners performed poorest. Moreover, Dutch listeners correctly recognized 'joy' by 76% while Taiwanese and Japanese listeners only identified it by 24% and 20%. These results show that perception of vocal emotion in an unknown language is both universal and language or culture specific. Thompson and Balkwill (2006) had twenty English-speaking listeners judge the emotive intent of utterances produced by English, German, Chinese, Japanese, and Tagalog speakers. Identification accuracy was above chance for all emotions expressed in all languages. Across languages, sadness and anger were more accurately recognized than joy and fear. The (English) listeners showed an in-group advantage for decoding emotional prosody, with highest recognition rates for

English utterances and lowest rates for Japanese and Chinese utterances. It would also indicate that emotional prosody is decoded by a combination of universal and culture-specific cues.

In addition, a few studies investigated the perception of vocal emotion by L2 learners. Some of them also studied the correlation between the learner's ability of recognizing emotions in the L2 and his/her L2 proficiency. For example, Graham et al. (2001) examined the ability of native and non-native speakers of English to identify emotions being portrayed by English speakers. They concluded that the ability to accurately identify emotions being portrayed through vocal cues in a second language (L2) may not be acquired by L2 learners without extensive exposure in a native context or without special attention to developing these skills in an instructional context. Moreover, an analysis of judgments made by learners of English as a Second Language (ESL) at different proficiency levels did not show an increase in ability to judge the emotional content of English speech with increased language proficiency. Chen (2005) studied how L2 English learners and L2 Dutch learners perceive emotional prosody in English and Dutch. She found that L1-transfer is an important strategy in interpreting pitch variation in L2. However, L2 learners may also activate their knowledge of intonational universals embodied in the biological codes. L2 learners at different levels seem to have acquired different degrees of understanding of the differences between their L1 and L2 and adjust their interpretation of pitch variation in L2 accordingly, with the advanced L2 learners being more successful than the beginning and the intermediate ones. Shochi, Gagnié, Rilliard, Erickson and Aubergé (2010) investigated the differences in the perception of six culturally encoded French social affects through audio and visual channels for French native listeners, naïve Japanese listeners and advanced Japanese learners of French. They found out that facial information cues seem to be more salient than auditory cues. The advanced Japanese learners of French recognized the emotions better than the naïve Japanese listeners; however, culture-specific attitudes (i.e. suspicious irony and obviousness) were confused by Japanese listeners (including advanced learners of French). This finding is in line with the conclusions of Van Bezooijen (1984) and Thompson and Balkwill (2006).

Ross et al. (1986) have shown that there is less use of short-term changes in F0 to express emotion in tone languages (in which short-term F0 contours are used to carry lexical information) than in Indo-European languages (in which F0 plays no lexical role). Thus it seems that in some cases at least, use of a particular acoustic feature in spoken language limits its use for the communication of emotion. At this point, I would like to propose that the prosodic space which languages may use is finite. The parameters (or dimensions) of the phonetic space (and the prosodic space within it) can be used to express linguistic as well as paralinguistic contrasts. This hypothesis holds that one can use a particular parameter in the phonetic space only once. In other words, if a language uses a prosodic parameter for linguistic purposes, it can no longer use the same parameter for non-/paralinguistic uses or – in a less extreme version of the hypothesis – cannot use the same parameter as effectively for the expression of paralinguistic or extralinguistic meanings. The prediction follows that speakers of a lexical tone language (such as Mandarin) have less room to express emotion through prosody (specifically through paralinguistic use of speech melody) than speakers of a non-tone language (such as Dutch or English). As a consequence, listeners of a tonal language will be less

intent on (and in fact less experienced in) decoding the paralinguistic use of prosody than listeners of a non-tonal language. Therefore, I conducted the present study (1) to test whether the prediction made here will apply to perception of Chinese emotional prosody by natives and non-natives. Chinese native listeners, then, should have more difficulty identifying emotion from speech melody, so that the in-group advantage may be reduced or even absent for them as compared with Dutch listeners; (2) to investigate to what extent (i) native, (ii) naïve non-native and (iii) advanced L2 learners of Chinese can perceive Mandarin emotions expressed vocally and also to find out what would be the differences between native and non-native listeners of Chinese in perceiving Chinese emotional prosody; (3) to find out whether advanced L2 learners of Chinese will perform better than naïve listeners in perceiving Chinese emotional prosody.

In this study six Chinese emotional prosodies have been studied by using the discrete-emotion approach: 'neutrality', 'happiness', 'anger', 'surprise', 'sadness' and 'sarcasm'.[8]

## 3.2 Methods

### 3.2.1 Participants

Twenty native Mandarin listeners (10 males, 10 females, mean age = 24 years), twenty naïve Dutch listeners (10 males, 10 females, mean age = 33 years) and twenty advanced Dutch learners of Chinese (10 males, 10 females, mean age = 20 years) voluntarily participated in the perception test. The Chinese listeners were bachelor and master students at the University of Science and Technology Beijing who hailed from different parts of China. The naïve Dutch listeners were mainly bachelor students of the Humanities Faculty of Leiden University in the Netherlands and volunteers with variable education backgrounds. None of the naïve Dutch listeners spoke any Mandarin. The advanced Dutch learners of Chinese were mainly third-year BA students in the Chinese Program of Leiden University; the others were one MA student and three outstanding second-year BA students. Early bilinguals were excluded; none of the students started learning Mandarin before they enrolled at Leiden University, i.e. after the age of eighteen. There was no special course designed for training these students to recognize emotions in Chinese in the curriculum.

### 3.2.2 Materials and procedure

In this study I used six Mandarin statements (e.g. *She is three months pregnant; He has been to Xiao Ge's place once*). The reasons that I selected these six sentences particularly are: (1) these six sentences contain all the tones of Mandarin, i.e. 'high-level tone', 'rising tone', 'falling-rising tone', 'falling tone' and 'neutral tone' (e.g. Howie 1976); (2) according to the consensus of the speakers, these sentences are semantically neutral but can easily be

[8] The discrete-approach here means that listeners have to choose emotions from a limited set of discrete emotion labels in a perception experiment. This is different from the so-called emotion tracking technique, which asks (trained) listeners to adjust a pointing device (e.g. the Feeltrace device, cf. Cowie et al. 2000) in a two-dimensional space with continuously variable axes in accordance with the perceived strength of emotional parameters.

expressed with different emotions; (3) both short and longer sentences were included, in case utterance length might play a role in the perception of emotional prosody. Each of the six statements was expressed in six different emotions (neutrality, happiness, anger, surprise, sadness and sarcasm).

Four native Mandarin speakers (2 males, 2 females, mean age = 45 years) whose mother tongue was standard Mandarin, voluntarily participated in the recording of the stimuli for the perception experiment. The stimuli were digitally recorded (44.1 KHz, 16 bits) in a sound-proofed booth through a Logitech external microphone. This procedure resulted in a stimulus set that consisted of 6 Chinese statements × 4 Mandarin speakers × 6 emotions = 144 discrete emotional utterances. The complete stimulus list can be found in Table 3.1.

All the participants including native Chinese listeners, naïve Dutch listeners and advanced learners of Chinese were asked to make a forced choice of the speaker's intended emotion, from the six given emotions immediately after they heard a stimulus. They also gave a confidence rating to each choice they made. A three-level confidence rating scale was used, with the following interpretation: 3 = 'The speaker expressed the intended emotion well. I am very confident in my answer', 2 = 'The speaker reasonably expressed the intended emotion. But I am not so sure about my answer' and 1 = 'The speaker did not express the intended emotion well. I made the choice only by guessing.' This confidence scale was introduced as a potential weighting factor. It would enable us to see which emotional utterances were identified by the listeners with more confidence and which were not. Therefore, we would be able to later correct the recognition rates based on the weighting. The entire experiment lasted 25 minutes, including the time for the participants to read the instructions in their native language before they started the test and a 6-second pause in between the emotional sentences for the participants to make a choice.

Each participant did the experiment individually in the presence of the experimenter. The stimuli were presented to the subject over closed headphones (but remained inaudible to the experimenter).

Table 3.1. *Stimulus list in Chinese (in Chinese characters), Pinyin Chinese-phonetic notation (including tone) and English glosses.* Macron '¯' = high-level tone, acute accent '´'= rising tone, haček 'ˇ' = falling-rising tone, grave accent '`' = falling tone; a syllable without tone mark has neutral tone.

| | |
|---|---|
| 1. | 是你。<br>shì nǐ<br>It is you. |
| 2. | 谢谢你。<br>xièxiè nǐ<br>Thank you. |
| 3. | 小王完全不知道这件事。<br>xiǎo wáng wánquán bù zhīdào zhè jiàn shì<br>Xiao Wang did not know about this matter. |
| 4. | 今天下午他不能来参加这个会。<br>jīntiān xiàwǔ tā bùnéng lái cānjiā zhège huì<br>He cannot attend the meeting this afternoon. |
| 5. | 她怀孕 3 个月了。<br>tā huáiyùn sān gè yuèle<br>She is three months pregnant. |
| 6. | 他去过小葛家一次。<br>tā qùguò xiǎo gé jiā yì cì<br>He has been to Xiao Ge's place once. |

### 3.3 Results

At the beginning of the present study, I explained the reason for introducing a confidence ratings scale, which would be used as a potential weighting factor. However, it turned out that there was no effect of weighting on the results, according to the statistical analysis. Therefore, I report unweighted identification results in this article only. The confidence ratings will be presented and analysed separately from the identification results. Table 3.2 shows the mean identification rates (in %) broken down by intended emotion. We can see from the table that native Chinese listeners and naïve Dutch listeners fell into very different confusion categories of emotions. For instance, Chinese listeners tended to mistake 'neutrality' for 'happiness' by 34.8%, while naïve Dutch listeners misidentified 'happiness' as 'anger' by 35.7%. Interestingly, the confusion categories that advanced Dutch learners of Chinese fell into are quite similar to those of naïve Dutch listeners. For example, Dutch learners of Chinese showed the same tendency as naïve Dutch listeners for 'happiness', which they often misidentified as 'anger' (30.2%). Moreover, both advanced Dutch learners of Chinese and naïve Dutch listeners misidentified 'sarcasm' as 'surprise' by 20%. This finding supports Chen's (2005) conclusion that L1-transfer is an important strategy in interpreting pitch variation in L2. 'Anger' was identified by the three listeners groups equally well, from which we conclude that anger could be the real basic and universal emotion to all human beings.

Table 3.2. *Confusion matrix of intended and perceived emotions by Chinese (upper panel), naïve Dutch (middle panel) listeners and advanced Dutch learners of Chinese (lower panel). Correct responses are located on the main diagonal (bold and shaded).*

| Intended | Ang | Hap | Neu | Sar | Sad | Spr |
|---|---|---|---|---|---|---|
| | Responded emotion by Chinese native listeners | | | | | |
| Angry | **56.3** | 4.8 | 10.2 | 5.2 | 10.0 | 13.5 |
| Happy | 12.1 | **37.3** | 34.8 | 1.7 | 0.8 | 13.3 |
| Neutral | 7.3 | 7.3 | **73.5** | 2.5 | 4.2 | 5.2 |
| Sarcastic | 11.7 | 17.5 | 34.0 | **17.3** | 3.1 | 16.5 |
| Sad | 13.3 | 8.1 | 32.7 | 4.4 | **37.1** | 4.4 |
| Surprised | 12.9 | 4.0 | 10.6 | 13.3 | 5.0 | **54.2** |
| | Responded emotion by Naïve Dutch listeners | | | | | |
| Angry | **52.6** | 4.0 | 15.1 | 9.5 | 7.9 | 10.9 |
| Happy | 35.7 | **20.4** | 14.1 | 5.8 | 3.6 | 20.4 |
| Neutral | 4.0 | 4.2 | **71.2** | 9.9 | 7.3 | 3.4 |
| Sarcastic | 13.3 | 6.5 | 15.7 | **28.8** | 16.1 | 19.6 |
| Sad | 5.2 | 2.4 | 28.6 | 10.5 | **49.2** | 4.2 |
| Surprised | 9.9 | 12.3 | 5.0 | 6.3 | 15.1 | **51.4** |
| | Responded emotion by advanced Dutch learners of Chinese | | | | | |
| Angry | **53.3** | 2.7 | 16.5 | 10.0 | 5.0 | 12.5 |
| Happy | 30.2 | **25.2** | 14.2 | 3.5 | 2.1 | 24.8 |
| Neutral | 5.2 | .8 | **80.2** | 2.9 | 9.6 | 1.3 |
| Sarcastic | 11.9 | 8.3 | 17.3 | **31.9** | 10.4 | 20.2 |
| Sad | 5.8 | 1.3 | 19.6 | 6.7 | **65.4** | 1.3 |
| Surprised | 6.0 | 9.4 | 2.5 | 10.0 | 3.8 | **68.3** |

Figure 3.1A shows the mean percentage of correct identification by the six intended emotions along the X-axis and broken down further by the three listener groups (in the legend). Figure 3.1B displays the same information, broken down first by listener group and second by emotion. Table 3.2 and Figure 3.1A-B together indicate that native Chinese, naïve Dutch listeners and advanced Dutch learners of Chinese were able to recognize discrete Chinese emotional prosodies above chance level (mean recognition rate: 48.6%, chance level: 16.7%). Moreover, emotions were identified much better than chance by each of the three listener groups. Even the Dutch naïve listeners obtained a score of 45.6% correct, closely followed by the native Chinese listeners (45.9% correct), and with the best performance obtained by the advanced Dutch learners of Mandarin (54.1% correct).

The data were analyzed by a repeated measures Analysis of Variance, with speaker, sentence and intended emotion as within-subjects factors and with listener type as a between-subjects factor. The dependent variable was the percentage of correctly identified emotions. Huynh-Feldt corrected degrees of freedom were used when the

assumption of sphericity was unreasonable. The difference between the three listener groups is statistically significant $F(2, 57) = 5.8$, $p = .005$, $p\eta^2 = .17$. A Bonferroni post-hoc test ($\alpha = .05$) showed that the advanced Dutch learner group performed better than the other two groups, which did not differ from each other. All main effects and all possible interactions were significant.

Surprisingly, native Chinese and naïve Dutch listeners followed a rather similar recognition order, such that they both found 'neutrality' the easiest emotion to identify, followed by 'anger', 'surprise', 'sadness', 'happiness' or 'sarcasm'. The detailed recognition order of the six emotional prosodies by the three listener groups is shown in Table 3.3.

Table 3.3. *Recognition order of the six emotional prosodies by native Chinese, Dutch naïve listeners and advanced Dutch learners of Chinese.**

| Listener group | Recognition order of the six emotional prosodies | | | | | |
|---|---|---|---|---|---|---|
| Native Chinese | neutrality > | anger > | surprise > | happiness > | sadness > | sarcasm |
| Dutch naïve | neutrality > | anger > | surprise > | sadness > | sarcasm > | happiness |
| Advanced learners | neutrality > | surprise > | sadness > | anger > | sarcasm > | happiness |

*: '>' means 'better identified than'.

The three listener groups all found 'neutrality' the easiest emotion to identify. This result confirms earlier findings that neutral prosody is identified more accurately than emotional prosody (Cornew et al. 2010). Moreover, correct identification rates of native Chinese and naïve Dutch listeners are strongly correlated ($r = .837$, $N = 6$, $p < .001$), showing again that native and non-native listeners of Mandarin display very similar cognitive behavior in identifying Chinese emotional prosody. In other words, the emotions that native listeners found easier to identify are also considered easier by naïve non-native listeners, and vice versa.

In Figure 3.1B, though advanced Dutch learners of Chinese identified neutrality most successfully in the six intended emotions, they actually followed a slightly different recognition order, indicating that the emotions which native Chinese and naïve Dutch listeners found difficult to identify are not necessarily difficult for advanced Dutch learners of Chinese to recognize (e.g. sadness and surprise). This finding supports Chen's (2005) study that L2 learners at different levels seem to have acquired different degrees of understanding of the differences between their L1 and L2, and adjust their interpretation of pitch variation in the L2 accordingly. Specifically, the advanced Dutch learners of Chinese showed higher identification rates for the emotions of 'sadness', 'surprise' and 'neutrality'. However, there was no significant difference between native Chinese listeners and naïve Dutch listeners, meaning that Chinese native listeners are not able to recognize emotions in their native language more successfully than naïve Dutch listeners. Interestingly, both naïve Dutch listeners and advanced Dutch learners

of Chinese showed better identification rates for 'sarcasm', which were 11.5% and 14.6% higher than that obtained by the Chinese listeners, respectively.

The findings of the perception experiment contradict the results demonstrated in previous studies that native listeners should recognize emotional prosody more accurately in their own language than non-native listeners do. (e.g. Dromey et al. 2004, Graham et al. 2001, Van Bezooijen 1984).

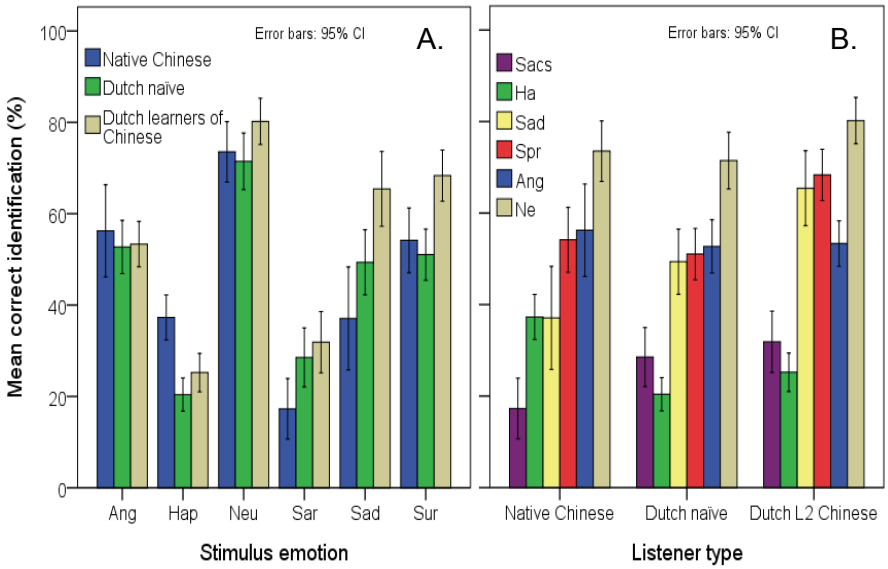



Figure 3.1A-B. *Percent correct identification of six intended emotions by native Chinese, naïve Dutch listeners and advanced Dutch learners of Chinese. Intended emotions in panel B are in ascending order of correct overall recognition. Confidence limits were computed for each bar on the basis of 20 listener means.*

Although there was no effect of weighting on the results for confidence, I would like to make use of the confidence ratings all the same to investigate the social behaviour of the listener groups. In this case, I just present mean confidence ratings and observe unexpected differences between the groups.

Figure 3.2 shows that Chinese native listeners were less confident (mean = 1.49) in their identifications than the Dutch listeners. Within the Dutch listeners the advanced learners of Mandarin were more confident (mean = 2.29) than the naïve listeners (mean = 1.96). The effect of listener group is significant by an ANOVA (same design as above), $F(2, 57) = 45.4$ ($p < .001$, $p\eta^2 = .614$). Bonferroni post-hoc tests revealed that the differences between all three groups were significant ($\alpha = .05$).

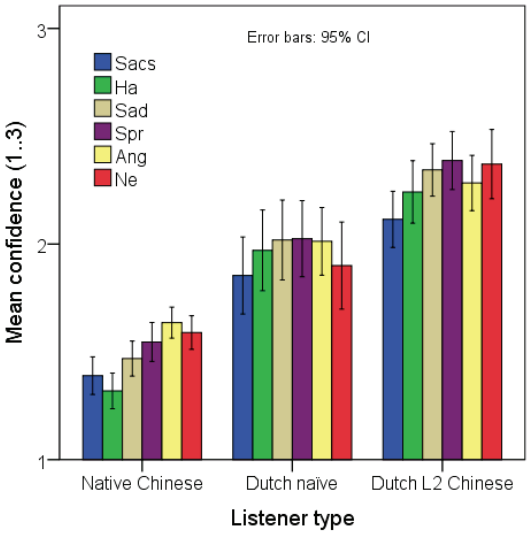



Figure 3.2. *Confidence rating (3 = most) of six intended emotions by native Chinese listeners, naïve Dutch listeners and advanced Dutch learners of Chinese. Intended emotions are ordered as in Figure 3.1B. Confidence limits were computed for each bar on the basis of 20 listener means.*

### 3.4 Conclusions and discussion

The results of this investigation indicate that Chinese native listeners are not able to identify emotions in their native language more accurately and confidently than naïve Dutch listeners and advanced Dutch learners of Chinese. Surprisingly, advanced Dutch learners of Chinese can recognize emotional prosody in Chinese significantly better than natives. This finding contradicts the conclusion of Graham et al.'s (2001) study that the ability to accurately identify emotions being portrayed through vocal cues in a second language may not be acquired by L2 learners without extensive exposure to such emotions in a native context or without special attention to developing these skills in an instructional context. Moreover, advanced Dutch learners of Chinese can identify Chinese emotional prosody substantially better than naïve Dutch listeners. This finding is in line with the result of Shochi et al.'s (2010) study that trained second language learners may recognize emotional prosody in the target language better than naïve ones.

Naïve non-native listeners can recognize unknown emotional prosody as well as natives do. Both non-native listener types (including the advanced learners of Chinese) can even identify some emotions (e.g. sadness, sarcasm) more successfully than native listeners. The in-group advantage found by other researchers, therefore, does not apply universally to all cultural groups (e.g. Chinese). Natives and naïve non-natives may have drawn on very similar cognitive resources when identifying emotional prosody, but the

incorrectly recognized emotional prosodies of natives and naïve non-natives may fall into different confusion categories. Advanced learners of Chinese followed a slightly different recognition order, indicating that emotions which are difficult for native and naïve non-native listeners to identify are not necessarily difficulty for them to recognize, for example: sadness and surprise. These findings are in line with the conclusions of Chen's (2005) study: L1-transfer is an important strategy in interpreting pitch variation in L2. L2 learners at different levels seem to have acquired different degrees of understanding of the differences between their L1 and L2 and adjust their interpretation of pitch variation in L2 accordingly. Therefore, the advanced learners of Chinese followed a slightly different recognition order of identifying emotions in Chinese while at the same time they fell into very similar confusion categories to the ones which the naïve Dutch listeners also fell into.

'Neutrality' is identified most accurately by all the listener groups in this investigation, which finding is in line with previous literature (Cornew et al. 2010). However, one might claim that 'neutrality' is the default response category so that its correct identification rate is predictably higher than that of other emotions. 'Anger' is recognized by the three listeners groups equally well, which means that it could possibly be a truly basic and universal emotion for all human beings. According to Darwin's evolution theory, 'anger' is supposed to signal aggression of the offended and to warn the offender to expect an aggressive reaction. In other words, 'anger' symbolizes danger to both the offender and the offended. Therefore, this emotion should be recognized equally well by all human beings regardless linguistic and cultural backgrounds, as people/animals have an instinct to sense, and protect themselves from, danger. From this point, we could further assume that emotion perception can be both universal and culture-specific depending on the particular emotions.

There may be several possible explanations for the finding that Chinese emotions were identified more successfully by (advanced) Dutch learners of Mandarin than by native Chinese listeners themselves.

1) Although voice quality and temporary changes in articulatory setting may also contribute to the expression of emotion (see introduction), these prosodic effects will not likely play a role in the comparison of Dutch and Mandarin.[9] Therefore, we would like to conclude that the prediction made at the beginning of the article is supported by our results. If a language – such as Mandarin – uses prosody for linguistic purposes, it can hardly use prosody for non-/paralinguistic uses. Since Mandarin uses prosody for linguistic purposes, specifically for expressing lexical tone contrast, Chinese listeners are less intent on the paralinguistic use of prosody than listeners of a non-tonal language (e.g. Dutch). In other words, listeners of a non-tonal language are generally better at recognizing emotional prosody than listeners of a tonal language. This explains why naïve Dutch listeners can recognize Chinese emotional prosody as well as natives, and why advanced Dutch L2 learners of Chinese can identify the same emotions even better. It is worth rerunning this

[9] The four Chinese native speakers portrayed the emotions in a normal voice. There was no special voice quality or phonation type, such as, hoarseness, roughness or smiling.

experiment with different linguistic groups to see if the results are similar, for example: British naïve listeners and British L2 learners of Chinese; or German naïve listeners and German learners of Chinese. The ultimate test of this explanation would be to test emotion recognition by speakers of another tone language, e.g. Vietnamese or Thai.

2) It may be the case that Chinese listeners normally recognize emotions by lexical or syntactic markers and contextual connotations (Xing 1999). If so, they are not experienced at identifying emotion through the audio channel only.[10]

3) Chinese society is quite reserved when it comes to the overt expression of emotions, either in speech or in other modes of communication (Klineberg 1938). Showing emotion in public is interpreted as a sign of weakness in China (Wu & Tseng 1985). Possibly, native listeners did not perform as well as advanced Dutch learners of Chinese because the latter have not been exposed to an emotion-free culture extensively, but the former are. Therefore, advanced Dutch learners of Chinese may be able to pick up some Chinese subtle emotions (e.g. sadness and sarcasm) more successfully than natives.

A basic problem with explanations of the type described above is that one can always reverse the argument. All human beings express emotions and are able to recognize emotions and respond to them. Now, if a cultural – or a linguistic code – prevents the speaker from expressing emotions plainly and overtly, the receiver (listener) is forced to attend to subtle expression of emotions. So on the one hand, such persons may be less sensitive to emotions as they have been exposed less to clear exemplars of the various affects, but on the other hand they may have learnt to be more attentive to subtle expression of emotions. In order to know what the effect is of growing up in an emotion-suppressive culture or linguistic environment, one can only turn to empirical observation.

[10] Lexical markers here refer to final particles in Chinese which can carry emotional information. Examples would be *ya* (friendly) or *a* (enthusiastic). Syntactic markers may be used to imply negative emotions such as the 'annoyance' marking construction *nán dào … (ma)?*, which is a rhetorical question confronting the listener with his/her ignorance (less negative with *ma* than without).

# Chapter Four

# Perception of Chinese Emotional Prosody Produced by Dutch Learners and Native Speakers of Chinese

**Abstract**

This chapter investigated the differences between perception of six Chinese emotional prosodies (neutrality, happiness, anger, surprise, sadness and sarcasm) produced by Dutch L2 speakers of Chinese and those encoded by native Chinese speakers (control group).[11] Twenty Chinese native listeners, 20 naïve non-native listeners (Dutch) and 20 advanced Dutch L2 learners of Chinese participated in each of the perception experiments. The results showed that the three listener groups recognized emotional prosodies encoded by Chinese natives significantly better than those produced by L2 speakers of Chinese. Also, the naïve non-native listeners could recognize the emotions in the unknown language as well as the natives did. Chinese native listeners, therefore, did not show an in-group advantage (i.e., identifying emotions in Chinese more accurately). Moreover, advanced Dutch L2 learners of Chinese could recognize native-produced Chinese emotional prosody significantly better than the Chinese native listeners themselves. A functional view is confirmed, which claims that listeners of a tonal language will be less intent on the paralinguistic use of prosody than listeners of a non-tonal language. Furthermore, it seems that, in some cases at least, the linguistic use of a particular acoustic feature in spoken language limits its use for the communication of emotion.

[11] This chapter will appear as Y. Zhu (2013b) Perception of Chinese emotional prosody produced by Dutch learners and native speakers of Chinese. *Chinese as a Second Language Research*.

## 4.1 Introduction

Perception and production of emotion is an essential part in human/animal communication (Darwin 1872). Scherer (2000) claimed that emotion needs to be viewed as a multicomponent phenomenon that should be studied simultaneously from biological, cognitive, physiological, cultural and linguistic perspectives. Each of these aspects may contribute to the shaping of emotions, and affect the way in which emotions are expressed and perceived within and across cultures. In this paper I will study the perception of emotions that are expressed vocally, concentrating on the question how successfully emotions produced by speakers of one language are perceived by listeners of a different language. Earlier findings obtained in such cross-cultural and/or cross-linguistic studies have borne out that the perception of emotion is partly universal and partly language/culture-specific. Some emotions, such as 'anger', 'sadness', 'neutrality', are produced and perceived through universal means of expression, meaning that these emotions are generally recognizable even by different cultural groups. To the extent that the vocal expression of emotions depends on general biological and physiological mechanisms shared by all humans, some emotions are distinguished by specific properties that are shared across languages and cultures. Non-native listeners will be able to recognize these emotions even if they are expressed by speakers of another language. However, for some emotions, for example 'sarcasm', 'disgust' or 'shame', may well be expressed in different ways depending on the native language and culture of the speaker, and may therefore not be successfully identified by listeners from a different linguistic or cultural background. As a case in point, Van Bezooijen (1984) studied ten emotional prosodies: neutral, disgust, surprise, shame, interest, joy, fear, contempt, sad, and angry. Her study aimed to find out how (Taiwanese) Chinese and Japanese listeners without any knowledge of Dutch, perceived the Dutch emotional prosodies. All three listener groups recognized the Dutch emotional prosodies well above chance level, with scores of 66, 37 and 33% correct for Dutch, Taiwanese and Japanese listeners, respectively. The Asian listeners' identifications correlated at $r = .6$ with the Dutch identification percentages but correlated somewhat more strongly between Japanese and Taiwanese ($r = .7$). Each of the emotions was identified better by the native listeners than by the Asian listeners, which points to a strong and consistent in-group advantage. Yet, the native and non-native identifications were relatively close together for 'sadness', 'fear', 'surprise' and 'anger' (< 30 percentage points difference) whilst other Dutch emotions were identified quite poorly: e.g. 'joy' and 'shame' (both 22% correct against 76 and 61% correct for the native listeners). We assume that the communication of first group of vocal emotions very much relies on a universal code whereas the latter two depend largely on language-specific cues. Thompson and Balkwill (2006) conducted a different experiment in which 20 English-speaking listeners judged the emotive intent of utterances spoken by male and female speakers of English, German, Chinese, Japanese, and Tagalog. Identification accuracy was above chance for all emotions expressed in all languages. Across languages, 'sadness' and 'anger' were more accurately recognized than 'joy' and 'fear'. The (English) listeners showed an in-group advantage for decoding emotional prosody, with highest recognition rates for English utterances and lowest rates for Japanese and Chinese utterances. It would also indicate that, again, emotional prosody is decoded by a combination of universal and culture-specific cues. Pell et al. (2009) carried out a similar study, in which they compared how monolingual speakers of

Argentine Spanish recognize basic emotions from pseudo-utterances ('nonsense speech') produced in their native language and in three foreign languages (English, German, Arabic). Results indicated that vocal expressions of basic emotions could be decoded in each language condition at accuracy levels exceeding chance, although Spanish listeners performed significantly better overall in their native language ('in-group advantage'). These findings suggest that the ability to understand vocally-expressed emotions in speech is partly independent of linguistic ability and involves universal principles, although this ability is also shaped by linguistic and cultural variables.[12]

In addition, a few previous studies investigated the perception of native produced vocal emotion by L2 learners. And some of them also studied the correlation between the learner's ability to recognize emotions in the L2 and his/her L2 proficiency. For example, Graham et al. (2001) examined the ability of native and non-native speakers of English to identify emotions being portrayed by English speakers. They concluded that the ability to accurately identify emotions being portrayed through vocal cues in a second language (L2) may not be acquired by L2 learners without extensive exposure in a native context or without special attention to developing these skills in an instructional context. Moreover, an analysis of judgments made by learners of English as a Second Language (ESL) at different proficiency levels did not show an increase in ability to judge the emotional content of English speech with increased language proficiency. Chen (2005) studied how L2 English learners and L2 Dutch learners perceive emotional prosody in English and Dutch. She found that L1-transfer is an important strategy in interpreting pitch variation in L2. However, L2 learners may also activate their knowledge of intonational universals embodied in the biological codes. L2 learners at different levels seem to have acquired different degrees of understanding of the differences between their L1 and L2 and adjust their interpretation of pitch variation in L2 accordingly, with advanced L2 learners being more successful than beginning and intermediate learners. Shoshi and Gagnié (2010) investigated the differences in the perception of six culturally encoded French social affects through audio and visual channels for French native listeners, naïve Japanese listeners and trained Japanese learners of French. The trained Japanese learners of French recognized the emotions better than the naïve Japanese listeners; however, culture-specific attitudes (i.e. 'suspicious irony' and 'obviousness') were confused by Japanese listeners (including trained listeners). Facial information cues seemed to be more salient than auditory cues.

However, previous studies on perception of vocal emotion by different cultural groups mainly concentrated on how native and non-native listeners perceived emotion produced by native speakers. There was little attention for perception of emotions encoded by L2 speakers, especially encoded by L2 speakers of a tonal language (e.g. Mandarin). Moreover, previous studies on vocal communication between native and non-native speakers of Chinese have mainly been carried out in the area of perception

[12] Pell et al. (2009) report a significant in-group advantage but omitted the responses to one of the emotions ('neutral'). However, when the Pell et al. data are aggregated over all six emotional categories, there is no significant in-group advantage for the Argentinean Spanish listeners.

or production of Mandarin lexical tones by L2 learners of Chinese (Flege 1997, Gandour 1983, Leather 1990, Stagray & Downs 1993, Wang et al. 1999). Therefore, the present study has the following aims:

(1) Investigate to what extent (i) native, (ii) naïve non-native and (iii) advanced second-language learners of Chinese can perceive Mandarin emotions encoded vocally by L2 speakers and also to find out what would be the differences between these listener groups in perceiving Chinese emotion vocally produced by native speakers.

(2) Test whether an in-group advantage really exists, which means native listeners should get a significantly higher recognition rate than non-natives.

In order to avoid terminological inconsistency I only use the term 'emotional prosody' in this chapter, and use it to refer to both vocally produced emotions (e.g. happiness, sadness, anger, fear, disgust) and attitudes (e.g. sincerity, irony, sarcasm). In this study six Chinese emotional prosodies have been studied by using the discrete-emotion approach: 'neutrality', 'happiness', (hot) 'anger', 'surprise', 'sadness' and 'sarcasm'.

## 4.2 Methods

Two perception experiments were conducted. The first perception experiment aimed to test how native Chinese listeners, naïve Dutch listeners and advanced Dutch learners of Chinese perceive Chinese emotional prosody produced by native Chinese speakers. The second perception experiment was designed to test how well the three listener groups recognize the same Chinese emotional prosodies when encoded by Dutch L2 speakers of Chinese.

### 4.2.1 Speakers

Four native Chinese speakers (2 males, 2 females, mean age = 45 years) whose mother tongue was standard Mandarin voluntarily took part in the recording of the stimuli for the first perception experiment. Four Dutch L2 speakers of Chinese (2 males, 2 females, mean age = 33 years) voluntarily participated in the recording of the stimuli for the second perception experiment. These four Dutch L2 speakers of Chinese, whose mother tongue was Dutch, were teachers in the Chinese department of Leiden University in the Netherlands. None of them were early bilinguals. They had learnt Chinese for 6 to 10 years, and they had been teaching Chinese for 2 to 10 years when the recording was made. All spent at least one year living or studying in mainland China or Taiwan.

### 4.2.2 Listeners

Twenty native Mandarin listeners (10 males, 10 females, mean age = 24 years), 20 naïve Dutch listeners (10 males, 10 females, mean age = 33 years) and 20 advanced Dutch learners of Chinese (10 males, 10 females, mean age = 20 years) voluntarily participated

in each of the perception experiments. The Chinese listeners were bachelor and master students at the University of Science and Technology Beijing who hailed from different parts of China. The naïve Dutch listeners were mainly bachelor students at the Humanities Faculty of Leiden University in the Netherlands and volunteers with variable education backgrounds. None of the naïve Dutch listeners spoke any Mandarin. The advanced Dutch learners of Chinese were mainly third-year BA students in the Chinese Program of Leiden University; the others were MA students and some outstanding second-year BA students. Early bilinguals were excluded; therefore, all students had learnt Mandarin after the age of puberty. There was no special course in the curriculum designed for training these students to recognize emotions in Chinese.

### 4.2.3 Materials and procedures

The first perception test used six Mandarin statements as vocal stimuli (e.g. *She is three months pregnant; He has been to Xiao Ge's place once*). Some of the sentences may be associated more readily with some emotions than with others but on aggregate the lexico-syntactic materials will not be biased towards specific emotions. Generally, speakers find it easier to pronounce meaningful sentences with specific emotions than they do with meaningless materials. This method has been widely used by other researchers in the vocal emotion study (e.g. Banse & Scherer 1996, Li et al. 2009, Van Bezooijen 1984, You et al. 2005, Zhang et al. 2006). We did not resort to the recording of meaningless *pseudo-utterances* (which has been proposed as an alternative solution by e.g. Castro & Lima 2010, Pell et al. 2009, Scherer et al. 1991) as these would be too difficult for L2 speakers of Chinese to vocally produce. The list of stimulus sentences is shown in Table 4.1.

Table 4.1. *Stimulus list in Chinese (Pinyin) with English glosses.*

| | |
|---|---|
| 1. | **Shì nǐ.*<br>'It is you.' |
| 2. | *Xièxiè nǐ.*<br>'Thank you.' |
| 3. | *Xiǎo wáng wánquán bù zhīdào zhè jiàn shì.*<br>'Xiao Wang does not know about this matter.' |
| 4. | *Jīntiān xiàwǔ tā bùnéng lái cānjiā zhège huì.*<br>'He cannot attend the meeting this afternoon.' |
| 5. | *Tā huáiyùn sān ge yuèle.*<br>'She is three months pregnant. ' |
| 6. | **Tā qùguò xiǎo gě jiā yì cì.*<br>'He has been to Xiao Ge's place once.' |

Note: '*' means sentences were excluded in the second perception experiment. Macron '¯' = high-level tone, acute accent '´'= rising tone, haček 'ˇ' = falling-rising tone, grave accent '`' = falling tone; a syllable without tone mark has neutral tone.

Each of the six statements was expressed in six different emotions – neutrality, happiness, (hot) anger, surprise, sadness and sarcasm – by the four native Chinese speakers. The stimuli were digitally recorded (44.1 KHz, 16 bits) in a sound-proofed booth through a Logitech desk-top microphone. This procedure resulted in a stimulus set that consisted of 6 Chinese statements × 4 Mandarin speakers × 6 emotions = 144 discrete emotional utterances.

For the second perception experiment, the four Dutch L2 speakers of Chinese were asked to express the same six emotional prosodies in Chinese. The stimuli were digitally recorded under the same conditions as in the first perception experiment. Two sentences were discarded from the stimulus set (see Table 4.1), as these two sentences were less well perceived by the three listener groups in the first perception test. Therefore, the final stimulus set for the second perception experiment consisted of 4 Chinese statements × 4 Dutch L2 speakers × 6 emotions = 96 discrete emotional utterances. It made the second experiment shorter than the first one. In the comparison between the two experiments, I only used the shared materials.

In both perception experiments, all the participants including native Chinese listeners, naïve Dutch listeners and advanced learners of Chinese were asked to make a forced choice of the speaker's intended emotion, from the six given emotions, immediately after they heard a stimulus. They also gave a confidence rating to each choice they made. A three-level confidence rating scale was used, with the following interpretation: 3 = 'The speaker expressed the intended emotion well. I am very confident in my answer', 2 = 'The speaker reasonably expressed the intended emotion. But I am not so sure about my answer' and 1 = 'The speaker did not express the intended emotion well. I made the choice only by guessing.' The confidence scale was introduced in order to obtain a potential weighting factor such that responses given with greater confidence would be weighted more heavily than responses that were largely based on guessing. The first experiment lasted 25 minutes and the second one lasted 15 minutes, including the time for the participants to read the instructions in their native language before they started the test and a 6-second pause in between the emotional utterances for the participants to make a choice.

Each participant did the experiment individually in the presence of the experimenter. The stimuli were presented to the subject over closed headphones (but remained inaudible to the experimenter).

## 4.3 Results

### 4.3.1 Identification of emotions

The results proved insensitive to any weighting based on response confidence. Therefore, I report unweighted identification results only. Tables 4.2 (which repeats Table 3.2) and 4.3 are confusion matrices of intended versus perceived emotions in the two perception experiments by the three listener groups, i.e., native Chinese listeners, Dutch naïve listeners and advanced Dutch learners of Chinese.

Table 4.2 (= Table 3.2). *Perception of emotional prosody produced by native Chinese speakers: Confusion matrix of intended and perceived emotions by Chinese (upper panel), naïve Dutch (middle panel) listeners and advanced Dutch learners of Chinese (lower panel). Correct responses are located on the main diagonal (shaded).*

| Intended | Ang | Hap | Neu | Sar | Sad | Spr |
|---|---|---|---|---|---|---|
| | Responded emotion by Chinese native listeners | | | | | |
| Angry | **56.3** | 4.8 | 10.2 | 5.2 | 10.0 | 13.5 |
| Happy | 12.1 | **37.3** | 34.8 | 1.7 | .8 | 13.3 |
| Neutral | 7.3 | 7.3 | **73.5** | 2.5 | 4.2 | 5.2 |
| Sarcastic | 11.7 | 17.5 | 34.0 | **17.3** | 3.1 | 16.5 |
| Sad | 13.3 | 8.1 | 32.7 | 4.4 | **37.1** | 4.4 |
| Surprised | 12.9 | 4.0 | 10.6 | 13.3 | 5.0 | **54.2** |
| | Responded emotion by Naïve Dutch listeners | | | | | |
| Angry | **52.6** | 4.0 | 15.1 | 9.5 | 7.9 | 10.9 |
| Happy | 35.7 | **20.4** | 14.1 | 5.8 | 3.6 | 20.4 |
| Neutral | 4.0 | 4.2 | **71.2** | 9.9 | 7.3 | 3.4 |
| Sarcastic | 13.3 | 6.5 | 15.7 | **28.8** | 16.1 | 19.6 |
| Sad | 5.2 | 2.4 | 28.6 | 10.5 | **49.2** | 4.2 |
| Surprised | 9.9 | 12.3 | 5.0 | 6.3 | 15.1 | **51.4** |
| | Responded emotion by advanced Dutch learners of Chinese | | | | | |
| Angry | **53.3** | 2.7 | 16.5 | 10.0 | 5.0 | 12.5 |
| Happy | 30.2 | **25.2** | 14.2 | 3.5 | 2.1 | 24.8 |
| Neutral | 5.2 | .8 | **80.2** | 2.9 | 9.6 | 1.3 |
| Sarcastic | 11.9 | 8.3 | 17.3 | **31.9** | 10.4 | 20.2 |
| Sad | 5.8 | 1.3 | 19.6 | 6.7 | **65.4** | 1.3 |
| Surprised | 6.0 | 9.4 | 2.5 | 10.0 | 3.8 | **68.3** |

Table 4.3. *Perception of emotional prosody produced by advanced Dutch L2 speakers of Chinese: Confusion matrix of intended and perceived emotions by Chinese (upper panel), naïve Dutch (middle panel) listeners and advanced Dutch learners of Chinese (lower panel). Correct responses are located on the main diagonal (shaded).*

| Intended | Ang | Hap | Neu | Sar | Sad | Spr |
|---|---|---|---|---|---|---|
| | Responded emotion by Chinese native listeners | | | | | |
| Angry | **25.6** | 6.3 | 34.4 | 12.8 | 8.1 | 12.8 |
| Happy | 3.4 | **37.8** | 21.3 | 12.2 | 3.1 | 22.2 |
| Neutral | 2.5 | 8.4 | **63.1** | 3.8 | 18.8 | 3.4 |
| Sarcastic | 13.1 | 15.9 | 27.2 | **21.3** | 11.9 | 10.6 |
| Sad | 8.4 | 2.5 | 27.8 | 5.3 | **47.2** | 8.8 |
| Surprised | 7.8 | 19.4 | 16.6 | 9.7 | 8.1 | **38.4** |
| | Responded emotion by Naïve Dutch listeners | | | | | |
| Angry | **38.4** | 4.7 | 14.4 | 17.2 | 12.5 | 12.8 |
| Happy | 14.1 | **29.4** | 12.8 | 11.3 | 8.4 | 24.1 |
| Neutral | 5.0 | 4.7 | **60.0** | 10.3 | 16.6 | 3.4 |
| Sarcastic | 13.8 | 15.6 | 18.4 | **19.1** | 14.7 | 18.4 |
| Sad | 6.9 | 1.9 | 25.9 | 9.7 | **42.2** | 13.4 |
| Surprised | 7.5 | 20.6 | 14.1 | 9.4 | 11.9 | **36.6** |
| | Responded emotion by advanced Dutch learners of Chinese | | | | | |
| Angry | **33.8** | 8.1 | 23.4 | 13.8 | 8.4 | 12.5 |
| Happy | 6.3 | **33.1** | 17.5 | 12.8 | 3.8 | 26.6 |
| Neutral | 2.8 | 6.3 | **59.1** | 5.3 | 24.7 | 1.9 |
| Sarcastic | 9.4 | 16.3 | 19.7 | **22.5** | 12.5 | 19.7 |
| Sad | 5.9 | 3.4 | 22.8 | 6.6 | **51.6** | 9.7 |
| Surprised | 5.0 | 17.8 | 18.1 | 7.8 | 7.8 | **43.4** |

The confusion matrices show that native Chinese, Dutch naïve listeners and advanced Dutch learners of Chinese perceived the six Chinese emotional prosodies produced by native Chinese speakers (mean recognition rate: 48.4%) substantially better than those encoded by Dutch L2 speakers of Chinese (mean recognition rate: 39.0%). For the perception of native-produced emotional prosody show quite different confusion patterns in identifying native produced emotional prosody. For instance, Chinese listeners tended to mistake 'happiness' mainly for 'neutrality' (34.8%) while naïve Dutch listeners massively misidentified 'happiness' as 'anger' (35.7%). In the perception of non-native emotional prosody, native Chinese listeners and naïve Dutch listeners showed a surprisingly similar confusion structure for the six emotions. For example, both native Chinese and Dutch naïve listeners strongly confused 'happiness' with 'surprise' (22.2% and 24.1%, respectively). Moreover, native Chinese and Dutch naïve listeners showed the same tendency of confusing 'sarcasm' with 'neutrality'.

In the perception of native-produced Chinese emotional prosody, advanced Dutch learners of Chinese performed significantly better than the other two listener groups. However, there was no significantly better listener group in the perception of L2 produced Chinese emotional prosody. Moreover, in both of the perceptual experiments, the confusion categories which advanced Dutch learners of Chinese fell into are quite similar to those of naïve Dutch listeners. For example, in the perception of native-produced emotional prosody, naïve Dutch listeners and advanced Dutch learners of Chinese both mistook 'anger' for 'neutrality' by 15.1% and 16.5%, respectively, and for 'surprise' by 10.9% and 12.5%, respectively. They both misrecognized 'happiness' as 'anger' by 35.7% and 30.2%. Furthermore, in perceiving non-native-produced emotional prosody, advanced Dutch learners of Chinese showed the exact same tendency as naïve Dutch listeners for 'sarcasm': they often confused 'sarcasm' with 'neutrality' (19.7%) and 'surprise' (19.7%); and naïve Dutch listeners confused it with 'neutrality' (18.4%) and 'surprise' (18.4%). In addition, in the second perception experiment the two Dutch listener groups both dramatically misidentified 'happiness' as 'neutrality' and 'surprise'; and they also confused 'neutrality' with 'sadness'. These findings support Chen's (2005) conclusion that L1-transfer is an important strategy in interpreting paralinguistic intonational meaning (e.g. emotional prosody) in L2.

Figure 4.1A-B shows the percent correct identification of six intended emotions by native Chinese, naïve Dutch listeners and advanced Dutch learners of Chinese in the two perception experiments. Figure 4.1A presents the results of the three listener groups perceiving emotional prosody produced by native Chinese speakers. Figure 4.1B shows the results of the three listener groups recognizing emotional prosody encoded by Dutch L2 speakers of Chinese.

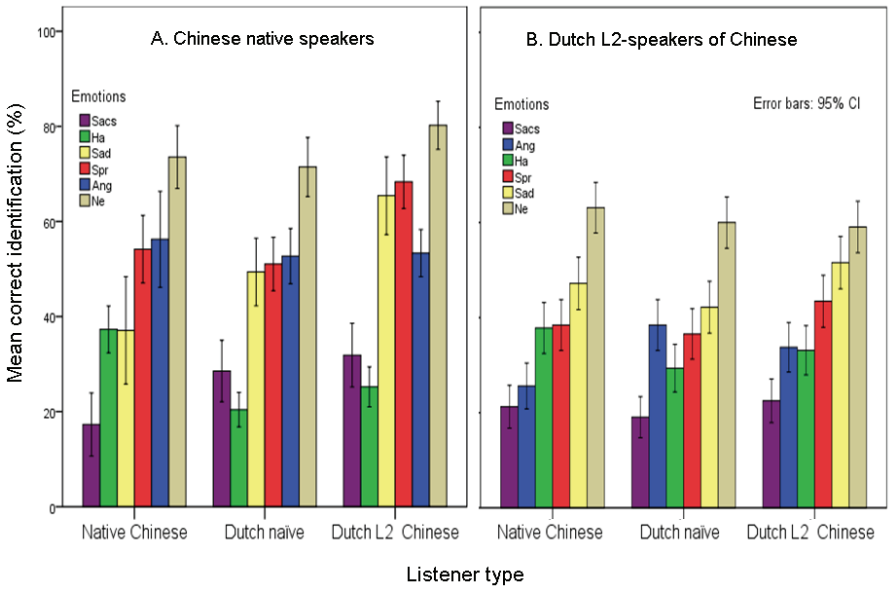



Figure 4.1A-B. *Percent correct identification of six intended emotions by native Chinese, naïve Dutch listeners and advanced Dutch learners of Chinese. Panel A presents the perceptual results of emotional prosody produced by native Chinese speakers (exp. 1). Panel B presents the perceptual results of emotional prosody encoded by advanced Dutch L2 speakers of Chinese (exp. 2). In each cluster of bars the intended emotions are ordered from left to right in ascending overall correct recognition rate. The order is indicated from top to bottom in the legends of the panels; note that the order differs between panels. The 95% confidence limits were computed for each bar on 20 listener means. Panel A repeats Figure 3.1B.*

Table 4.2, Table 4.3 and Figure 4.1A-B together indicate that native Chinese, naïve Dutch listeners and advanced Dutch learners of Chinese were able to recognize discrete Chinese emotional prosody, whether produced by native or by non-native Chinese speakers, above chance level (mean recognition rates: 48.6% and 39.0%, chance level: 16.7%). Moreover, emotions were identified (much) better than chance by each of the three listener groups in the two perception tests. In the perception of native-produced emotional prosodies, even the Dutch naïve listeners obtained a score of 45.6% correct, closely followed by the native Chinese listeners (45.9% correct), and with the best performance obtained by the advanced Dutch learners of Mandarin (54.1% correct). Furthermore, the difference between the three listener groups is statistically significant by a one-way Analysis of Variance, $F(2, 57) = 5.8$, $p = .005$. A Bonferroni post-hoc test ($\alpha = .05$) showed that the advanced Dutch learner group performed better than the other two groups in perception of native-produced emotional prosody. The other two listener groups did not differ from each other. In the perception of non-native-produced Chinese emotional prosody, there was no statistical significance between the three listener groups, even though advanced Dutch learners of Chinese performed slightly better than the other two groups (2% or 3% higher). This indicates that native Chinese, naïve Dutch listeners and advanced Dutch learners of Chinese performed

equally well/poorly in perceiving Chinese emotional prosody encoded by L2 speakers of Chinese.

Somewhat surprisingly, the success with which native Chinese listeners and Dutch naïve listeners identified vocal emotions in each of the perception experiments was correlated. Emotions that native listeners found difficult (or easy) to identify were also found difficult (or easy) for naïve listeners. For example, both groups identified 'anger', 'surprise', and 'sadness' more successfully than 'happiness' and 'sarcasm' in the first perceptual experiment. However, they found 'sadness' and 'surprise' less difficult to recognize than 'anger' in the second perceptual experiment. In Figure 4.1A, the order of difficulty among the six emotions was somewhat different for the advanced Dutch learners of Chinese than for the other listener groups. Specifically, the advanced Dutch learners of Chinese showed much higher identification rates for the emotions of 'sadness', 'surprise' and 'neutrality' portrayed by the native speakers. This finding supports Chen's study (2005) that L2 learners at different levels seem to have acquired different degrees of understanding of the differences between their L1 and L2, and adjust their interpretation of pitch variation in L2 accordingly.

### 4.3.2 Confidence rating

In the second part of this results section, I will analyze the confidence ratings. Although, as mentioned earlier in this section, there was no effect of weighting on the results and only unweighted identification results were presented, I would like to make use of the confidence ratings all the same to investigate the social behaviour of the listener groups. In this case, I just present means and observe unexpected differences between the groups.

Figure 4.2A-B shows the confidence rating of six intended emotions by native Chinese listeners, naïve Dutch listeners and advanced Dutch learners of Chinese in the two perceptual experiments. Figure 4.2A shows that Chinese native listeners were less confident than the non-native listeners (mean = 1.49) in their identifications of native-produced emotions. Within the Dutch listeners, the advanced learners of Mandarin were more confident (mean = 2.29) than the naive listeners (mean = 1.96). The effect of listener group is significant by a one-way ANOVA, $F(2, 57) = 45.4$, ($p < .001$). Bonferroni post-hoc tests revealed that all differences between the three groups were significant ($\alpha = .05$). Figure 4.2B shows the opposite result that Chinese native listeners were as confident as advanced Dutch learners of Chinese (mean = 2.30), but Dutch naïve listeners were the least confident (mean = 1.92). Therefore, it can be concluded that native Chinese listeners are more confident in identifying emotional prosody produced by (Dutch) non-native speakers. The reason for this behavior is not clear since one would expect listeners to be more confident when having to make decisions based on materials produced by speakers who share the same linguistic code.

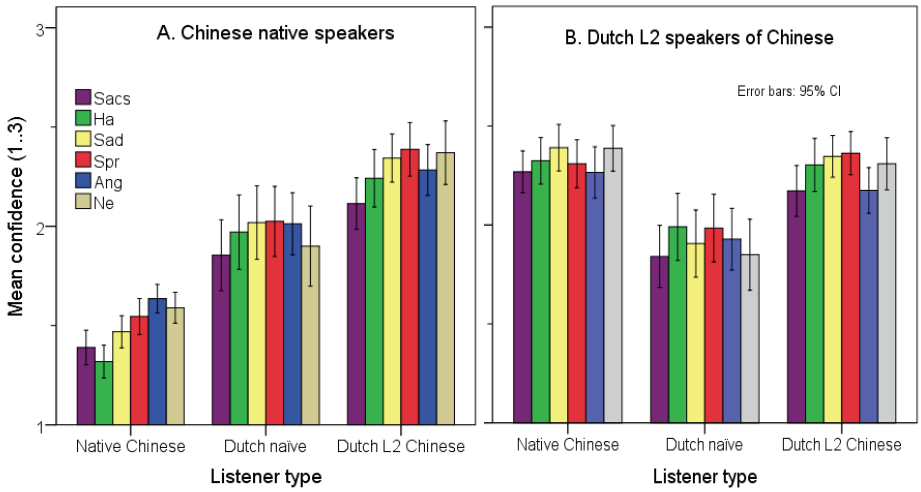



Figure 4.2A-B: *Confidence rating (3 = most) of six intended emotions by native Chinese listeners, naïve Dutch listeners and advanced Dutch learners of Chinese. Panel A presents the perceptual results of emotional prosody produced by native Chinese speakers. Panel B presents the perceptual results of emotional prosody encoded by advanced Dutch L2 speakers of Chinese. In each cluster in both panels the intended emotions are ordered as in Figure 4.1A. Confidence limits are based on 20 listener means.*

### 4.3.3 Peripheral findings of the production of emotional prosody by L2 speakers[13]

In the last part of this section, I would like to address briefly the production of the Dutch L2 speakers who produced the stimuli for the present study. The results show that even advanced Dutch L2 speakers of Chinese are generally not as good as native speakers at vocally expressing emotions in Chinese. This could possibly be explained along the lines of Ross et al. (1986), who found that less use was made of short-term changes in F0 to express emotion in tone languages (in which short-term F0 contours are used to carry lexical information) than in Indo-European languages (in which F0 plays no lexical role). It would appear, then, that in some cases at least, linguistic use of a particular acoustic feature in spoken language limits its use for the communication of emotion. Obviously, native Chinese speakers can produce emotional prosody in Chinese without having problems of getting the lexical tones right. Therefore, we may assume that native speakers of a tonal language automatically encode lexical information when producing emotions in their native language. However, L2 speakers of Chinese may not know how to pronounce Chinese lexical tones correctly while at the same time expressing emotional prosody on top of the lexical tones. Perhaps, that is why the three listener groups did not perceive the non-native-produced emotional prosodies as well as those encoded by natives. It is unclear whether Dutch L2 speakers

[13] A more comprehensive account of the production of emotional prosody in L2 and L1 is presented in Chapters 5 and 6.

of Chinese used an L1-transfer strategy during their production of Chinese emotional prosody, meaning that they would use Dutch vocal cues to express emotions in Chinese. If they did, then the naïve Dutch listeners should have picked up the Dutch features straightaway. However, the Dutch naïve listeners did not show any better scores in this case. Or it might be that L2 speakers have used the L1-transfer strategy in producing emotion in their second language. But it may not have been easy for Dutch naïve listeners to pick up the cues, if they could not distinguish between Chinese lexical tones and emotional prosody. In other words, L1-transfer strategy in terms of production of emotional prosody might not work on L2 speakers of a tonal language. The viability of this speculation can only be checked when an acoustic analysis is applied. I will come back to this point in Chapter 6.

## 4.4 Conclusions and discussion

The results of this investigation indicate that native-produced emotional prosodies were recognized better by all the listener groups than those expressed by L2 speakers. In other words, emotional prosodies produced by L2 speakers of Chinese were less recognizable overall than those encoded by natives. Nevertheless, the three listener groups could recognize emotions well above chance level, regardless the speaker type. Moreover, the results showed that the three listener groups could recognize some negative emotions equally well, regardless the speaker type, for example, 'anger' and 'sadness'. More specifically, the native-produced 'anger' was recognized by the three listeners groups equally well in first perception experiment. These findings support previous studies which claimed that recognizing negative emotions from vocal cues, independently of language, is compatible with evolutionary views that vocal signals associated with threat that much must be highly salient to ensure human survival (Ohman et al. 2001, Tooby & Cosmides 1990). According to Darwin's evolution theory, 'anger' signals aggression towards the offended and warns the offender to expect an aggressive counter-reaction. In other words, 'anger' symbolizes danger to both the offender and the offended. Therefore, this emotion should be recognized equally well by all human beings irrespective of their linguistic and cultural backgrounds, as people/ animals have an instinct to sense, and protect themselves from, danger. Furthermore, 'neutrality' is identified most accurately by all the listener groups in this investigation, which finding is in line with previous literature (Cornew et al. 2010). However, one might claim that neutrality is the default response category so that its correct identification rate is higher than that of other emotions as a result of response uncertainty. Therefore, we can infer that emotion perception could be universal to some extent.

The recognition rates of 'happiness' and 'sarcasm' were relatively low in the two perception experiments. This finding is compatible with the previous finding that 'happiness' was recognized relatively poorly in Mandarin where the emotion must be recognized through audio channel only (e.g. Banse & Scherer 1996, Castro & Lima 2010, Liu & Pell 2012). And the recognition rates of 'happiness' by the three listener groups also varied regardless of speaker type. This implies that perception of positive emotions depends more on listener's linguistic and cultural background. Moreover,

even some primary emotions (e.g. 'anger') can be expressed in variable ways by different speaker groups. For instance, 'anger' encoded by the L2 speakers was identified poorly by all the listeners. It implies that, if an emotional prosody was not able to trigger the language-specific cues borne in L1/L2 listeners' linguistic knowledge, the recognition rate of the prosody should decrease to some extent. Altogether, we can conclude that perception of emotional prosody can be partly universal and partly language-or-culture specific. It means that, on one hand, native and non-native listeners (including L2 listeners) might have drawn on the universal resources embedded in human beings to decode some primary emotional prosodies, e.g. 'anger', 'sadness' or 'neutrality'. On the other hand, they might have also resorted to their own language-or-culture specific cues, which vary very much from culture to culture, when perceiving emotional prosody produced by different speaker groups. According to the results of this chapter, it seems that the non-native produced emotional prosodies neither properly triggered the universal cues nor the language-or-culture-specific cues of all the listener groups. This is why they were not recognized as well as those produced by the native speakers.

There are also some other findings in this chapter. In the perception of native-produced emotional prosodies, Chinese native listeners were not able to identify emotions more accurately and confidently than naïve Dutch listeners and advanced Dutch learners of Chinese. Surprisingly, advanced Dutch learners of Chinese recognized emotional prosody in Chinese significantly better than Chinese natives did themselves. This finding contradicts the conclusion of Graham's study (2001) that the ability to accurately identify emotions being portrayed through vocal cues in a second language may not be acquired by L2 learners without extensive exposure to such emotions in a native context or without special attention to developing these skills in an instructional context. Moreover, advanced Dutch learners of Chinese can identify Chinese emotional prosody substantially (and significantly) better than naïve Dutch listeners. Possible explanations for this finding can be found below. This finding confirms the result of Shoshi and Gagnié's study (2010) that trained second language learners may recognize emotional prosody in the target language better than listeners with no experience in the target language.

Moreover, naïve non-native listeners can recognize unknown emotional prosody as well/poorly as natives, regardless of speaker type. The in-group advantage found by other researchers therefore does not apply universally to all cultural groups (e.g. Chinese), which means that native listeners would perform significantly better than non-native listeners in perceiving emotional prosody in their L1. Natives and naïve non-natives may have drawn on very similar cognitive resources when identifying emotional prosody; even the incorrectly recognized emotional prosodies of natives and non-natives may fall into similar confusion categories. However, the detailed cognitive resources are still not known at the present stage. It might be that there are some universal cognitive resources shared by the two listener groups. Advanced learners of Chinese followed a slightly different order of success in the perception of native-produced emotional prosody. It indicates that emotions which are difficult for native and naïve non-native listeners to identify, are not necessarily difficult for them to recognize, for example: sadness and surprise. These findings support the conclusions of

Chen's study (2005) to some extent: L1-transfer is an important strategy in interpreting pitch variation in L2. L2 learners at different levels of proficiency seem to have acquired different degrees of understanding of the differences between their L1 and L2 and adjust their interpretation of pitch variation in L2 accordingly.

Finally, I will briefly summarize some additional findings that relate to the performance of the two speaker groups in the present study.

Firstly, L2 speakers are not able to vocally produce emotions in their L2 as well as natives, even though their Chinese proficiency is high. This finding supports previous studies (Gorelick & Ross 1987, Lieberman & Michaels 1962, Ross et al. 1986, Scherer et al. 1984): although spoken language constrains emotional expression to some extent, linguistic and emotional expression can be dissociated and typically function independently of one another. From this observation we can possibly conclude that a second language might constrain emotional expression more than a first language does, especially when the second language is a tonal language (e.g. Chinese).

Secondly, we do not know at this stage whether this L1-transfer strategy is also used by L2 speakers in production of emotional prosody in their L2, since the Dutch non-native listeners did not pick up any Dutch vocal cues from the Chinese emotional prosodies encoded by Dutch L2 speakers of Chinese. Otherwise, they would have scored better than the native listeners in the perceptual test.

There may be several possible explanations for the findings that (i) L2-produced emotional prosodies were overall less recognizable than those produced by natives, and (ii) Chinese emotions were identified more successfully by (advanced) Dutch learners of Mandarin than by native Chinese listeners themselves.

First of all, Ross et al. (1986) have shown there is less use of short-term changes in F0 to express emotion in tone languages (in which short-term F0 contours are used to carry lexical information) than in Indo-European languages (in which F0 plays no lexical role). Thus it seems that, in some cases at least, use of a particular acoustic feature in spoken language limits its use for the communication of emotion. This insight is incorporated into a functional view which claims that the prosodic space which languages may use is finite. The parameters (or dimensions) of the phonetic space (and of the prosodic space within it) can be used to express linguistic as well as paralinguistic contrasts. The functional principle holds that one can use a particular parameter in the phonetic space only once. It follows from the functional principle that if a language uses a prosodic parameter for linguistic purposes, it can no longer use the same parameter for non-/paralinguistic uses – or, in a less extreme version of the theory – cannot use the same parameter as effectively for the expression of paralinguistic or extralinguistic meanings. The prediction follows that speakers of a lexical tone language (such as Mandarin) have less room to express emotion through prosody (specifically through paralinguistic use of speech melody) than speakers of a non-tone language (such as Dutch or English). Apparently, native Chinese speakers can filter out Mandarin lexical tones automatically during the production of emotional

prosody in their native language, but L2 speakers of Chinese cannot. In this case, L2 speakers of Mandarin cannot easily separate emotional prosody from lexical tones during their production of Chinese emotional prosody, so that they cannot express it as well as natives. As a consequence of a functional view, listeners of a tonal language will be less intent on (and well in fact be less experienced in) decoding the paralinguistic use of prosody than listeners of a non-tonal language. In other words, listeners of a non-tonal language are generally better at recognizing emotional prosody than listeners of a tonal language. This would explain why naïve Dutch listeners can recognize Chinese emotional prosody as well as natives, and why advanced Dutch L2 learners of Chinese can identify the same emotions even better.

It is worthwhile rerunning this experiment with different linguistic groups to see if the results are similar, for example: British naïve listeners and British L2 learners of Chinese; or German naïve listeners and German learners of Chinese. The ultimate test of this explanation would be to examine emotion recognition by speakers of another tone language, e.g. Vietnamese or Thai. The prediction, obviously, would be that such listeners should recognize emotions in Mandarin more poorly than native Chinese listeners do – since (i) they are relatively insensitive to emotional prosody because their mother tongue is a tone language, and (ii) because being non-native listeners they are not familiar with the expression of emotion in the target language.

Secondly, the unexpected results might be caused by the absence of particles in the Chinese stimuli. In everyday Chinese speech particles often appear at the end of a sentence, carrying considerable emotional information that is alternatively expressed by intonation in other languages. Since this kind of lexical markers were deliberately left out in the present study due to the research purpose, their absence might have affected the perception of the emotional prosodies by L1 listeners but not by the Dutch listeners. Moreover, Dutch listeners might generally be more intent on the message in the sentence prosody, according to the functional view. That is possibly why the Chinese L1 listeners did not perform better than the non-native listeners in the perception experiments in which the stimuli with no final particles attached were presented through audio channel only. Testing this hypothesis is beyond the scope of the present study. Part of the endeavor would be to determine how much use Mandarin and Dutch make of particles expressing emotions on the part of the speaker and what the division of work would be between the use of such particles and emotional prosodies.

Thirdly, Chinese society is quite reserved when it comes to the overt expression of emotion, either in speech or in other modes of communication (Klineberg 1938). Showing emotion in public is interpreted as a sign of weakness in China (Wu & Tseng 1985). If this is indeed the case, then native speakers of Chinese will have had little exposure to clear instances of vocally expressed emotions. This would explain why the native Chinese listeners did relatively poorly when instructed to identify vocally expressed emotions in Chinese. It would also explain why Dutch listeners obtained equal or better identification rates for the Chinese emotions than the native listeners themselves. Especially the advanced Dutch learners of Chinese can pick up some

Chinese subtle emotions produced by native speakers (e.g. sadness and sarcasm) more successfully than natives.

Further studies could be carried out in the areas of second-language acquisition and cognitive psychology to find out more about the perception and production of emotional prosody by natives and non-natives/L2 learners of the target language.

# Chapter Five

# Production of Emotional Prosody in L2 and in L1

**Abstract**

This chapter investigated how well Dutch L2 speakers of Chinese produced the six emotional prosodies (neutrality, happiness, anger, surprise, sadness and sarcasm) in their L2 (Chinese) and how well they produced the same emotional prosodies in their L1 (Dutch).[14] Two recognition studies were carried out in this chapter. The first recognition study was designed to test how well Dutch L2 speakers of Chinese produced the six emotional prosodies in Chinese. Native Chinese speakers participated in the first recognition study as the control group. The second recognition study aimed to find out how well the same Dutch L2 speakers of Chinese expressed the same vocal emotions in their native language – Dutch. Twenty Chinese native listeners, 20 naïve listeners (Dutch), and 20 advanced Dutch L2 learners of Chinese participated in the first recognition study and another 20 Dutch native listeners participated in the second recognition study as listeners/judges. The results showed that emotional prosodies produced by L2 speakers of Chinese in their L2 were overall less recognizable than those encoded by Chinese natives. Dutch L2 speakers of Chinese are better at vocally producing emotions in their L1 than in the L2. The prediction made in the beginning of this chapter is confirmed, which claims that second language limits L2 speakers' communication of emotion. A detailed acoustic analysis of selected stimuli is deferred to Chapter 6. The results also show that the naïve Dutch listeners could recognize the emotions in the unknown language (Mandarin Chinese) as well as the natives did. Moreover, naïve Dutch native listeners showed an in-group advantage in that they identified the same emotions in Dutch more accurately than in Mandarin Chinese.

[14] This chapter is the first part of Y. Zhu (2013). Production of emotional prosody in L2 and in L1 (submitted).

## 5.1 Introduction

Perception and production of emotion is an essential part of human/animal communication (Darwin 1872). The research question 'can listeners infer emotion from vocal cues?' has been studied by many researchers (e.g. Frick 1985, Scherer 1986, Standke 1992, Van Bezooijen 1984). These studies all show that listeners are rather good at inferring affective state and speaker attitude from vocal expression. Furthermore, the previous studies also claim that the vocal expression of emotions is differentially patterned (Scherer 1996). There is considerable evidence that emotion produces changes in respiration, phonation and articulation. A large number of different emotional and motivational states are indexed and communicated by specific acoustic characteristics of the concurrent vocalizations (Scherer 1989). The acoustic variables that are strongly involved in the production of vocally expressed emotion are summarized in Scherer's (1991, 1996) studies. However, previous studies mainly touched on the vocal production of emotion by native speakers from one particular linguistic group but not on speakers' L2.

Another finding is from Ross et al. (1986). They have shown that there is less use of short-term changes in F0 to express emotion in tone languages (in which short-term F0 contours are used to carry lexical information) than in Indo-European languages (in which F0 typically plays no lexical role). Thus it seems that, in some cases at least, use of a particular acoustic feature in spoken language limits its use for the communication of emotion. Inspired by Ross et al. I would like to predict that, if a language uses a prosodic parameter for linguistic purposes, it will have less space for non-/para-linguistic uses of the same cue. If this prediction were true, it would effectively mean that speakers of a lexical tone language (such as Mandarin) have less room to express emotion through prosody (specifically through paralinguistic use of speech melody) than speakers of a non-tone language (such as Dutch or English).

Therefore, I carried out the present study to:

(1) Investigate: (i) how L2 speakers of Chinese vocally produce emotions in Chinese; and how they portray the same emotions in their L1; and what would be the differences; (ii) what would be the differences between Chinese native and L2 speakers of Chinese vocally producing emotion in Chinese; (iii) will a tonal language limit the vocal production of emotion?

(2) As a secondary aim, investigate to what extent (i) native, (ii) naïve non-native and (iii) advanced second-language learners of Chinese can perceive Mandarin emotions encoded vocally by L2 speakers and to find out how these listener groups perceive Chinese emotions vocally produced by native speakers.

There is little literature which studied the first research question properly, especially when the target language is a tonal language, such as Mandarin. Anolli et al. (2008) conducted research on vocal production of emotion by Chinese and Italian young adults. They confirm that different emotions may be expressed through variations in the modulation of vocal cues, in both cultures; on the other hand, differences in the specific patterns of vocal cues in expressing emotions were identified between Chinese and Italian participants. Fortunately, there are a few studies which touched on perception of emotional prosody by both native and non-native listeners. To some

extent, previous findings all claimed that perception of emotion by different culture groups is partly universal and partly language/culture-specific. For instance, Van Bezooijen (1984) studied ten emotional prosodies: neutral, disgust, surprise, shame, interest, joy, fear, contempt, sad, and angry. Her study aimed to find out how (Taiwanese) Chinese and Japanese listeners, who did not have any knowledge of Dutch, perceived Dutch emotional prosodies at the sentence level. Perceptual experiments showed that Dutch native listeners got the highest correct identification rate and Japanese listeners performed poorest. But both of the listener groups performed well above chance level. Graham et al. (2001) examined the ability of native and non-native speakers of English to identify emotions being portrayed by English speakers. They concluded that the ability to accurately identify emotions being portrayed through vocal cues in a second language (L2) may not be acquired by L2 learners without extensive exposure in a native context or without special attention to developing these skills in an instructional context. Moreover, an analysis of judgments made by learners of English as a Second Language (ESL) at different proficiency levels did not show an increase in ability to judge the emotional content of English speech with increased language proficiency. Thompson and Balkwill (2006) conducted a similar experiment in which 20 English-speaking listeners judged the emotive intent of utterances spoken by male and female speakers of English, German, Chinese, Japanese, and Tagalog. Identification accuracy was above chance for all emotions expressed in all languages. Across languages, 'sadness' and 'anger' were more accurately recognized than 'joy' and 'fear'. The (English) listeners showed an in-group advantage for decoding emotional prosody, with highest recognition rates for English utterances and lowest rates for Japanese and Chinese utterances. This would indicate that, again, emotional prosody is decoded by a combination of universal and culture-specific cues. Shoshi and Gagnié (2010) investigated differences in the perception of six culturally encoded French social affects through audio and visual channels for French native listeners, naïve Japanese listeners and trained Japanese learners of French. The trained Japanese learners of French recognized the emotions better than the naïve Japanese listeners did; however, culture-specific attitudes (i.e. 'suspicious irony' and 'obviousness') were confused by Japanese listeners (including trained listeners). Facial information cues seemed to be more salient than auditory cues.

This chapter will focus on the vocal production of emotion in speakers' L2 and L1. In order to avoid terminological inconsistency I only use the term 'emotional prosody' in this chapter, and use it to refer to both vocally produced emotions (e.g. happiness, sadness, anger, fear, disgust) and attitudes (e.g. sincerity, irony, sarcasm). In this study six Chinese emotional prosodies have been studied by using the discrete-emotion approach: neutrality, happiness, anger, surprise, sadness and sarcasm. [15] Sentences expressed in different emotions are used as stimuli for the perception experiment. There is no semantic link between the sentences.

[15] Discrete emotion theory assumes that humans universally express and recognize a small number (six to eight) of basic cross-culturally shared 'core' emotions, which are communicated through innate mechanisms (for a survey of positions see Ekman & Friesen 1971, Colombetti 2009; see also footnote 8).

## 5.2 Methods

Two recognition studies were conducted: the first recognition study aimed to test how well Dutch L2 Chinese speakers vocally expressed emotions in their second language, compared to Chinese native speakers (the control group). Actually, this recognition study is the combination of the first and the second judgment study presented in Chapters 3 and 4, respectively. It is now reviewed from the production perspective; the second recognition study was designed to test how well the same Dutch L2 speakers of Chinese vocally produced emotions in their mother tongue i.e. Dutch. Three groups of listeners who were used as judges voluntarily participated in the first recognition study; 20 native Dutch listeners were used as judges in the second recognition study.

### 5.2.1 Speakers

Four Dutch L2 speakers of Chinese (2 males, 2 females, mean age = 33 years) voluntarily participated in the recording of the stimuli for the two recognition studies. These four Dutch L2 speakers of Chinese, whose mother tongue was Dutch, were teachers from the Chinese department of Leiden University in the Netherlands. None of them were early bilinguals. They had learnt Chinese for 6 to 10 years; and they had been teaching Chinese for 2 to 10 years at the time the recordings were made. All had spent at least one year living or studying in mainland China or Taiwan. In order to set up a control group for the first recognition study, four native Chinese speakers (2 males, 2 females, mean age = 45 years) whose mother tongue was standard Mandarin, voluntarily took part in the recording of the stimuli for the perception experiment. The four Chinese speakers were amateur actors/actresses who all had stage performance experience.

### 5.2.2 Listeners

Twenty native Mandarin listeners (10 males, 10 females, mean age = 24 years), 20 naïve Dutch listeners (10 males, 10 females, mean age = 33 years) and 20 advanced Dutch learners of Chinese (10 males, 10 females, mean age = 20 years) voluntarily participated in the first recognition study. They were asked to decide which emotion was intended by the speaker and how confident they were of their choice. The results can tell us how well the speakers had produced the six emotions in their L1 and L2. The Chinese listeners were bachelor and master students at the University of Science and Technology Beijing, who hailed from different parts of China. The naïve Dutch listeners were mainly bachelor students at the Humanities Faculty at Leiden University in the Netherlands and volunteers with variable education backgrounds. None of the naïve Dutch listeners spoke any Mandarin. The advanced Dutch learners of Chinese were mainly third-year BA students in the Chinese Program of Leiden University; the others were MA students and some outstanding second-year BA students. Early bilinguals were excluded; therefore, all students had learnt Mandarin after the age of eighteen. There was no special course in the curriculum designed for training these students to recognize emotions in Chinese.

Twenty Dutch native listeners who did not have any Chinese knowledge voluntarily participated in the second recognition study as listeners. They were bachelor or master students at Leiden University, majoring in linguistics. Although they were different subjects from those naïve Dutch listeners who took part in the first recognition study, both groups represent the same population statistically.

### 5.2.3 Materials and procedures

#### 5.2.3.1 First recognition study

The first recognition study includes two perception experiments: the first perception experiment was set up to test how well the Chinese control group vocally expressed the six emotions in Chinese. In this experiment, I used six Mandarin statements (e.g. *She is three months pregnant; He has been to Xiao Ge's place once*). The reasons that I selected these six sentences particularly are: (1) these six sentences contain all the tones in Mandarin, i.e. 'high-level tone', 'rising tone', 'falling-rising tone', 'falling tone' and 'neutral tone' (e.g. Howie 1976); (2) according to the consensus of the speakers, these sentences are semantically neutral but can easily be expressed with different emotions; (3) both short and longer sentences were included, in case utterance length might play a role in the perception of emotional prosody. Each of the six statements was expressed in six different emotions (neutrality, happiness, anger, surprise, sadness and sarcasm).

The second perception experiment was carried out to find out how well the Dutch L2 speakers of Chinese encoded the six emotions in their second language, compared to the control group. In this experiment, two Mandarin statements were discarded from the original six Mandarin statements, as they were not very well perceived by the three groups of listeners in the first perception experiment. The list of stimulus sentences for the first recognition study is shown in Table 5.1.

In the first perception experiment, each of the six Mandarin statements was vocally expressed in six different emotions (neutrality, happiness, anger, surprise, sadness and sarcasm) by the four native Chinese speakers. The stimuli were digitally recorded (44.1 KHz, 16 bits) in a sound-proofed booth through a Logitech desk-top microphone. This procedure resulted in a stimulus set that consisted of 6 Chinese statements × 4 Mandarin speakers × 6 emotions = 144 discrete emotional utterances.

In the second perception experiment, the four Dutch L2 speakers of Chinese were asked to express the same six emotional prosodies in Chinese. The stimuli were digitally recorded under the same conditions as in the first perception experiment. Two sentences were discarded from the stimulus set (see Table 5.1), as these two sentences were less well perceived by the three listener groups in the first perception test. Therefore, the final stimulus set for the second perception experiment consisted of 4 Chinese statements × 4 Dutch L2 speakers × 6 emotions = 96 discrete emotional utterances. It made the second experiment shorter than the first one.

Table 5.1. *Stimulus list in Chinese (Pinyin orthography) with English glosses.*

| | |
|---|---|
| 1. | **Shì nǐ.*<br>'It is you.' |
| 2. | *Xièxiè nǐ.*<br>'Thank you.' |
| 3. | *Xiǎo wáng wánquán bù zhīdào zhè jiàn shì.*<br>'Xiao Wang does not know about this matter.' |
| 4. | *Jīntiān xiàwǔ tā bùnéng lái cānjiā zhège huì.*<br>'He cannot attend the meeting this afternoon.' |
| 5. | *Tā huáiyùn sān ge yuèle.*<br>'She is three months pregnant. ' |
| 6. | **Tā qùguò xiǎo gě jiā yì cì.*<br>'He has been to Xiao Ge's place once.' |

Note: '*' means sentence was excluded in the second perception experiment. Macron '¯' = high-level tone, acute accent '´'= rising tone, haček 'ˇ' = falling-rising tone, grave accent '`' = falling tone; a syllable without tone mark has neutral tone.

In both perception experiments, all the participants (native Chinese listeners, naïve Dutch listeners and advanced learners of Chinese) were asked to make a forced choice of the speaker's intended emotion, from the six given emotions, immediately after they heard a stimulus. They also gave a confidence rating to each choice they made. A three-level confidence rating scale was used, with the following interpretation: 3 = 'The speaker expressed the intended emotion well. I am very confident of my answer', 2 = 'The speaker reasonably expressed the intended emotion. But I am not so sure about my answer' and 1 = 'The speaker did not express the intended emotion well. I made the choice mainly by guessing.' This confidence scale was introduced as a potential weighting factor. It would enable us to see which emotional utterances were identified by the listeners with more confidence and which were not. Therefore, we would later be able to compute the recognition rates based on the weighting. The first experiment lasted 25 minutes and the second one lasted 15 minutes, including the time for the listeners to read the instructions (in their native language) before they started the experiment and a 6-second pause in between the emotional utterances for the listeners to make a choice.

Each participant did the experiment individually in the presence of the experimenter. The stimuli were presented to the subject over closed headphones (but remained inaudible to the experimenter).

#### 5.2.3.2 Second recognition study

The second recognition study only included one perception experiment, in which 20 native Dutch listeners perceived the six emotions produced by the same four Dutch L2 speakers of Chinese, but in their mother tongue, i.e. Dutch. In this experiment, the four Mandarin statements used in the first recognition study were translated into Dutch by

the four Dutch L2 speakers of Chinese where sentence length, syntactic structure, syllables and sentence meaning were well controlled. Therefore, the final stimulus set for the perception experiment consisted of 4 Dutch statements × 4 Dutch L2 speakers × 6 emotions = 96 discrete emotional utterances. The list of stimulus sentences for the second recognition study is shown in Table 5.2. The same procedure as in the first recognition study was used to obtain the judgments.

Table 5.2. *Stimulus list of Dutch sentences with broad IPA transcription and English glosses.*

| | |
|---|---|
| 1. | *Dank je wel.*<br>dɑŋk jə ʋɛl<br>'Thank you.' |
| 2. | *Xiaowang weet dat helemaal niet.*<br>ʃau ʋɑŋ ʋetɑt heləmal nit<br>'Xiao Wang does not know about this matter.' |
| 3. | *Vanmiddag kan hij niet naar de vergadering.*<br>ʋɑmɪdɑχ kɑni nit nar də ʋərɣadərɪŋ<br>'He cannot attend the meeting this afternoon.' |
| 4. | *Zij is drie maanden zwanger.*<br>zɛɪ ɪs dri mandə zʋɑŋər<br>'She is three months pregnant. ' |

## 5.3 Results

At the beginning of the present study, I explained the reason of introducing a confidence rating scale, which would be used as a potential weighting factor. However, it turned out that there was no effect of weighting on the results, according to the statistical analysis. Therefore, I report unweighted identification results in this chapter only.

### 5.3.1 Results of production

#### 5.3.1.1 Production of emotional prosody in speakers' L2

Tables 5.3 and 5.4 (repeated and extended versions of Tables 3.2 and 4.2, respectively) are confusion matrices of intended versus perceived emotions in the two perception experiments by the three listener groups, i.e., native Chinese listeners, Dutch naïve listeners and advanced Dutch learners of Chinese. The confusion matrices show that native Chinese, Dutch naïve listeners and advanced Dutch learners of Chinese perceived the six Chinese emotional prosodies produced by native Chinese speakers (overall recognition rate: 48.7%) substantially better than those encoded by Dutch L2 speakers of Chinese (overall recognition rate: 39.3%). Figure 5.1A (which repeats

Figures 3.1B and 4.1A) presents the results of the three listener groups perceiving emotional prosody produced by native Chinese speakers (the control group). Figure 5.1B (which repeats Figure 4.1B) shows the results of the three listener groups recognizing emotional prosody encoded by Dutch L2 speakers of Chinese.

Table 5.3. *Perception of Chinese emotional prosody produced by native Chinese speakers: Confusion matrix of intended and perceived emotions by Chinese (upper panel), naïve Dutch (middle panel) listeners and advanced Dutch learners of Chinese (lower panel). Correct responses are located on the main diagonal (shaded). This table repeats and extends Tables 3.2 and 4.2).**

| Intended | Responded emotion by Chinese native listeners | | | | | | | Grand mean |
|---|---|---|---|---|---|---|---|---|
| | Ang | Hap | Neu | Sar | Sad | Spr | Mean | |
| Angry | **56.3** | 4.8 | 10.2 | 5.2 | 10.0 | 13.5 | | |
| Happy | 12.1 | **37.3** | 34.8 | 1.7 | .8 | 13.3 | | |
| Neutral | 7.3 | 7.3 | **73.5** | 2.5 | 4.2 | 5.2 | 46.0 | |
| Sarcastic | 11.7 | 17.5 | 34.0 | **17.3** | 3.1 | 16.5 | | |
| Sad | 13.3 | 8.1 | 32.7 | 4.4 | **37.1** | 4.4 | | |
| Surprised | 12.9 | 4.0 | 10.6 | 13.3 | 5.0 | **54.2** | | |
| Total | 20.4 | 13.5 | 26.6 | 6.3 | 13.4 | 19.6 | 100 | |
| | Responded emotion by Naïve Dutch listeners | | | | | | | |
| | Ang | Hap | Neu | Sar | Sad | Spr | Mean | |
| Angry | **52.6** | 4.0 | 15.1 | 9.5 | 7.9 | 10.9 | | |
| Happy | 35.7 | **20.4** | 14.1 | 5.8 | 3.6 | 20.4 | | |
| Neutral | 4.0 | 4.2 | **71.2** | 9.9 | 7.3 | 3.4 | 46.0 | |
| Sarcastic | 13.3 | 6.5 | 15.7 | **28.8** | 16.1 | 19.6 | | 48.7 |
| Sad | 5.2 | 2.4 | 28.6 | 10.5 | **49.2** | 4.2 | | |
| Surprised | 9.9 | 12.3 | 5.0 | 6.3 | 15.1 | **51.4** | | |
| Total | 19.1 | 7.4 | 25.8 | 10.4 | 17.8 | 18.6 | 100 | |
| | Responded emotion by advanced Dutch learners of Chinese | | | | | | | |
| | Ang | Hap | Neu | Sar | Sad | Spr | Mean | |
| Angry | **53.3** | 2.7 | 16.5 | 10.0 | 5.0 | 12.5 | | |
| Happy | 30.2 | **25.2** | 14.2 | 3.5 | 2.1 | 24.8 | | |
| Neutral | 5.2 | .8 | **80.2** | 2.9 | 9.6 | 1.3 | 54.0 | |
| Sarcastic | 11.9 | 8.3 | 17.3 | **31.9** | 10.4 | 20.2 | | |
| Sad | 5.8 | 1.3 | 19.6 | 6.7 | **65.4** | 1.3 | | |
| Surprised | 6.0 | 9.4 | 2.5 | 10.0 | 3.8 | **68.3** | | |
| Total | 16.5 | 7.8 | 24.8 | 9.8 | 20.2 | 21.1 | 100 | |

*Note: 'Mean' = mean correct identification rate of each listener group; 'Grand mean' = mean correct identification rate of the three listener groups.

Table 5.4. *Perception of Chinese emotional prosody produced by Dutch L2 speakers of Chinese: Confusion matrix of intended and perceived emotions by Chinese (upper panel), naïve Dutch (middle panel) listeners and advanced Dutch learners of Chinese (lower panel). Correct responses are located on the main diagonal (shaded). This table repeats and extends Table 4.3).**

| Intended | Responded emotion by Chinese native listeners | | | | | | | Grand mean |
|---|---|---|---|---|---|---|---|---|
| | Ang | Hap | Neu | Sar | Sad | Spr | Mean | |
| Angry | **25.6** | 6.3 | 34.4 | 12.8 | 8.1 | 12.8 | | |
| Happy | 3.4 | **37.8** | 21.3 | 12.2 | 3.1 | 22.2 | | |
| Neutral | 2.5 | 8.4 | **63.1** | 3.8 | 18.8 | 3.4 | 39.0 | |
| Sarcastic | 13.1 | 15.9 | 27.2 | **21.3** | 11.9 | 10.6 | | |
| Sad | 8.4 | 2.5 | 27.8 | 5.3 | **47.2** | 8.8 | | |
| Surprised | 7.8 | 19.4 | 16.6 | 9.7 | 8.1 | **38.4** | | |
| Total | 10.9 | 16.2 | 27.0 | 9.1 | 20.2 | 16.4 | 100 | |
| | Responded emotion by Naïve Dutch listeners | | | | | | | |
| | Ang | Hap | Neu | Sar | Sad | Spr | Mean | |
| Angry | **38.4** | 4.7 | 14.4 | 17.2 | 12.5 | 12.8 | | |
| Happy | 14.1 | **29.4** | 12.8 | 11.3 | 8.4 | 24.1 | | |
| Neutral | 5.0 | 4.7 | **60.0** | 10.3 | 16.6 | 3.4 | 38.0 | |
| Sarcastic | 13.8 | 15.6 | 18.4 | **19.1** | 14.7 | 18.4 | | 39.3 |
| Sad | 6.9 | 1.9 | 25.9 | 9.7 | **42.2** | 13.4 | | |
| Surprised | 7.5 | 20.6 | 14.1 | 9.4 | 11.9 | **36.6** | | |
| Total | 16.8 | 12.9 | 26.3 | 8.4 | 18.5 | 16.1 | 100 | |
| | Responded emotion by advanced Dutch learners of Chinese | | | | | | | |
| | Ang | Hap | Neu | Sar | Sad | Spr | Mean | |
| Angry | **33.8** | 8.1 | 23.4 | 13.8 | 8.4 | 12.5 | | |
| Happy | 6.3 | **33.1** | 17.5 | 12.8 | 3.8 | 26.6 | | |
| Neutral | 2.8 | 6.3 | **59.1** | 5.3 | 24.7 | 1.9 | 41.0 | |
| Sarcastic | 9.4 | 16.3 | 19.7 | **22.5** | 12.5 | 19.7 | | |
| Sad | 5.9 | 3.4 | 22.8 | 6.6 | **51.6** | 9.7 | | |
| Surprised | 5.0 | 17.8 | 18.1 | 7.8 | 7.8 | **43.4** | | |
| Total | 13.7 | 13.5 | 24.0 | 9.1 | 21.0 | 17.6 | 100 | |

*Note: 'Mean' = mean correct identification rate of each listener group; 'Grand mean' = mean correct identification rate of the three listener groups.

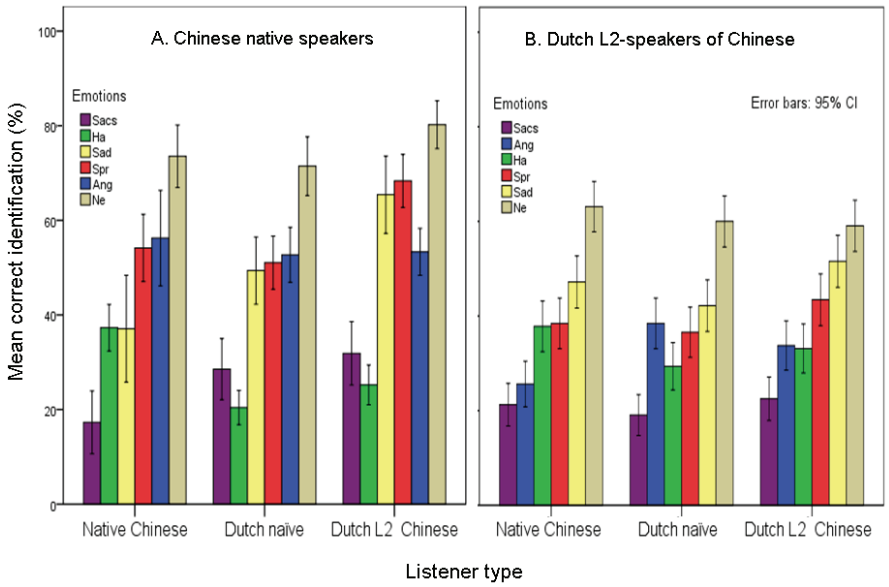



Figure 5.1A-B (= Figure 4.1A-B). *Percent correct identification of six intended Chinese emotions by native Chinese, naïve Dutch listeners and advanced Dutch learners of Chinese in the two perception experiments. Figure A presents the perceptual results of Chinese emotional prosody produced by native Chinese speakers. Figure B presents the perceptual results of Chinese emotional prosody encoded by Dutch L2 speakers of Chinese. The correct recognition rate of native Chinese speakers is 10 percentage points higher than that of Dutch L2 speakers of Chinese.*

Tables 5.3-4 and Figure 5.1A-B together indicate that both native and non-native produced Chinese emotional prosodies were recognized by the three listener groups above chance level (chance level: 16.7%). It means that Dutch L2 speakers of Chinese were able to vocally produce emotions in the L2. However, given the significant difference between the mean recognition rates of the three listener groups in the two perception experiments, we can conclude that Dutch L2 speakers of Chinese are generally not as good as native Chinese speakers are at vocally expressing emotions in Chinese. Specifically, the native-produced Chinese emotional prosodies (mean = 48.7%) were significantly better recognized overall than those produced by the L2 speakers (mean = 39.3%) by the three listener groups. This would mean that, in some cases at least, L2 speakers might not have the same ability as natives of expressing emotional prosody in the L2, even though their language proficiency of the L2 is very high. In the present study, native Chinese speakers can produce emotional prosody in Chinese without having problems of getting the lexical tones right. Therefore, we can assume that native speakers of a tonal language could automatically work out lexical information when producing emotions in their native language. However, L2 speakers of Chinese may not know how to pronounce Chinese lexical tones correctly while at the same time expressing emotional prosody on top of the lexical tones. In other words, even though Chinese lexical tones might limit the production of Chinese emotional

prosody to both L1 and L2 speakers, they might limit L2 speakers more. Perhaps, that is why the three listener groups did not perceive non-native-produced Chinese emotional prosodies as well as those encoded by Chinese natives.

#### 5.3.1.2 Production of emotional prosody in speakers' L1

The second recognition study only included one perception experiment in which 20 Dutch native listeners were used as judges to test how well the same four Dutch L2 speakers of Chinese had produced the six emotional prosodies in their L1. Table 5.5 is the confusion matrix of intended versus perceived emotions in the perception experiment by the Dutch native listeners. It shows that the emotional prosodies produced in Dutch by the same Dutch L2 speakers of Chinese were recognized by the native Dutch listeners above chance level (chance level = 16.7%). Moreover, the overall correct recognition rate of the Dutch native listeners increased dramatically from 39% when the emotional prosodies expressed in speakers' L2-Chinese to 57% when the emotional prosodies were produced in the speakers' mother tongue, i.e. Dutch. This indicates that the four Dutch L2 speakers of Chinese are able to express emotional prosodies both in the L2 and in their native language. However, they are better at producing emotional prosody in their L1 than producing it in their L2. It further supports the claim that L2 limits an L2 speaker's production of emotional prosody.

Table 5.5. *Perception of Dutch emotional prosody produced by Dutch L2 speakers of Chinese: Confusion matrix of intended and perceived emotions by native Dutch listeners. Correct responses are located on the main diagonal (shaded and bolded).*

| Intended | Responded emotion by Dutch native listeners | | | | | | |
|---|---|---|---|---|---|---|---|
| | Ang | Hap | Neu | Sar | Sad | Spr | Mean |
| Angry | **55.6** | 9.5 | 9.0 | 12.5 | 7.9 | 3.7 | |
| Happy | 2.3 | **44.9** | 4.6 | 17.4 | 1.2 | 12.5 | |
| Neutral | 15.9 | 13.0 | **68.3** | 17.6 | 19.7 | 5.3 | 57.0 |
| Sarcastic | 13.4 | 7.9 | 3.9 | **34.3** | 6.3 | 6.0 | |
| Sad | 8.1 | 9.7 | 13.4 | 4.4 | **64.6** | .0 | |
| Surprised | 5.6 | 15.0 | .7 | 13.9 | .5 | **72.5** | |
| Total | 16.3 | 13.1 | 20.0 | 10.0 | 18.9 | 21.2 | 100 |

### 5.3.2 Perception results

Although the focus of this chapter is the vocal production of emotion in speakers' L2 and L1, I would like to analyse the perceptual performance of the listener groups in the two recognition studies, as production can never be separated from perception in the study of speech, especially not in the study of emotional prosody. I believe that investigating the perception of the emotional prosodies can tell us more about the production of the emotions.

#### 5.3.2.1 Perception of native and non-native Chinese emotional prosodies

As can be seen from Table 5.3, for the perception of native-produced emotional prosody, native Chinese listeners and naïve Dutch listeners show quite different confusion patterns. For instance, Chinese listeners tended to mistake 'happiness' mainly for 'neutrality' (34.8%) while naïve Dutch listeners massively confused 'happiness' and 'anger' (35.7%). In the perception of non-native produced emotional prosody, native Chinese listeners and naïve Dutch listeners showed a surprisingly similar confusion structure for the six emotions. For example, both native Chinese and Dutch naïve listeners strongly confused 'happiness' with 'surprise' (22.2% and 24.1% respectively). Moreover, native Chinese and Dutch naïve listeners showed the same tendency of mistaking 'sarcasm' for 'neutrality'.

In the perception of native-produced emotional prosodies, even the Dutch naïve listeners obtained a score of 45.6% correct, closely followed by the native Chinese listeners (45.9% correct), and with the best performance obtained by the advanced Dutch learners of Mandarin (54.1% correct). The difference between the three listener groups is statistically significant by a one-way Analysis of Variance, $F(2, 57) = 5.8$, $p = .005$. A Bonferroni post-hoc test ($\alpha = .05$) showed that the advanced Dutch learner group performed better than the other two groups in perception of native-produced emotional prosody. The other two listener groups did not differ from each other. In the perception of non-native-produced Chinese emotional prosody, there were no statistically significant differences between the three listener groups, even though advanced Dutch learners of Chinese performed slightly better than the other two groups (2 or 3 percentage points higher). This indicates that native Chinese, naïve Dutch listeners and advanced Dutch learners of Chinese performed equally well/poorly in perceiving Chinese emotional prosody encoded by L2 speakers of Chinese.

In both of the perceptual experiments, the confusion categories which advanced Dutch learners of Chinese fell into, are quite similar to those of naïve Dutch listeners. For example, in perceiving non-native-produced emotional prosody, advanced Dutch learners of Chinese showed the exact same tendency as naïve Dutch listeners for 'sarcasm': they often misidentified 'sarcasm' as 'neutrality' (19.7%) and 'surprise' (19.7%); and naïve Dutch listeners mistook it for 'neutrality' (18.4%) and 'surprise' (18.4%). These observations suggest that L1-transfer is an important strategy in interpreting paralinguistic meaning (e.g. emotional prosody) in L2.

#### 5.3.2.2 Perception of the Dutch emotional prosodies by the Dutch listeners

Some confusion tendencies shown in Table 5.4 can be also seen in Table 5.5. For instance, the Dutch native listeners tended to mistake 'anger' mainly for 'sarcasm' (12.5%) when the emotion was produced in their L1; the naïve Dutch listeners also strongly misperceived 'anger' as 'sarcasm' (17.2%) when the emotion was expressed in Chinese (L2). Moreover, the same thing happened with perceiving 'sadness' and 'surprise': the naïve Dutch native listeners in both recognition studies confused 'sadness' mainly with 'neutrality', and confused 'surprise' mainly with 'happiness' regardless

the language in which the emotional prosodies were produced. From this, we can infer that both listeners and speakers use L1-transfer as a strategy in perception and production of emotional prosody in L2. Furthermore, we can also infer that knowing the meaning of the utterances does not influence the perception of emotional prosody very much. In other words, perception of emotional prosody is universal to some extent.

### 5.3.3 Combining the two recognition studies

In this section, I will report a summary analysis of the two recognition studies. Figure 5.2 presents the percentage of correctly identified emotions by seven combinations of speaker and listener type. Braces define speaker-listener combinations that do not differ significantly from each other (Bonferroni post-hoc test with $\alpha = .05$).

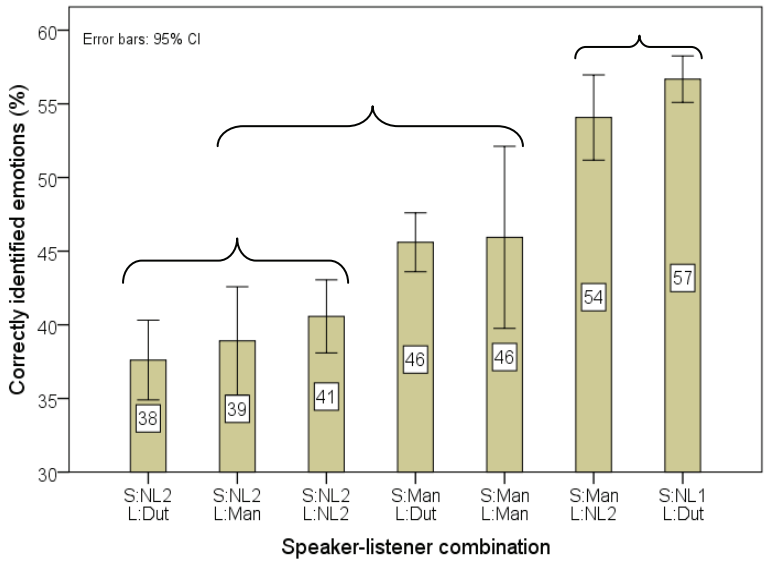



Figure 5.2. *Percent correct identification (%) of combined speaker and listener types.* Conditions under the same brace do not differ significantly from each other by a Bonferroni post hoc test ($\alpha = .05$).*

* Note: 'S'= speaker type; 'L' = listener type; 'Man'= native Mandarin speaker; 'Dut'= native Dutch listener (naïve); 'NL2' = emotional prosodies produced by Dutch L2 speakers of Chinese in their L2-Mandarin; 'NL1'= emotional prosodies produced by Dutch L2 speakers of Chinese in their L1, i.e. Dutch.

As can be seen from Figure 5.2, emotions, whether expressed in one's L1 or L2, are overall recognized above chance level, regardless listener type. It shows that perception and production of emotional prosody is universal to some extent. However, emotions vocally produced in the speaker's L1 are much more recognizable than those expressed in the speaker's L2. It indicates that, although the speaker is able to produce emotional prosody both in his L1 and L2, he produces emotional prosody in his L1 more

recognizably than in his L2, especially when the L2 is a tonal language. Also, native Dutch listeners identified emotions expressed in their L1 more successfully than those produced in Chinese. This finding supports the finding of Thompson and Balkwill's (2006) study that L1 listeners show an in-group advantage by decoding emotional prosody in their L1 more successfully than in other, non-native languages.

Surprisingly, advanced Dutch learners of Chinese recognized emotional prosodies produced by native Chinese speakers significantly better than native Chinese listeners and naïve Dutch listeners did. It indicates that advanced learners of a second language who have acquired high language proficiency of the L2 are better at interpreting paralinguistic meanings in the L2 than naïve listeners do. This finding is also compatible with Shoshi and Gagnié's (2010) finding that trained Japanese learners of French recognized the French emotions better than the naïve Japanese listeners.

## 5.4 Conclusions and discussion

The results of this chapter indicate that emotional prosodies produced by L2 speakers of Chinese were less recognizable overall than those encoded by natives. In other words, L2 speakers are not able to vocally produce emotions in their L2 as well as natives are, even though their Chinese proficiency is high. Furthermore, neither are they able to vocally portray emotions in their L2 as well as they doing in their L1. These finding confirm previous studies (Gorelick & Ross 1987, Lieberman & Michaels 1962, Ross et al. 1986, Scherer et al. 1984): spoken language constrains emotional expression to some extent; and these two systems can be dissociated and function independently of one another. From this point, we can possibly conclude that speaking in second language might constrain emotional expression more than first language does. However, the three listener groups could recognize emotion well above chance level, regardless the speaker type. Moreover, 'neutrality' is identified most accurately by all the listener groups in the present study, which finding is in line with previous literature (Cornew et al. 2010). Therefore, we can infer that emotion production is universal to some extent.

Native-produced 'anger' is recognized reasonably well in the two recognition studies, but 'anger' encoded in Chinese by the L2 speakers of Chinese is identified poorly by all the three listener groups. It seems that the L2 speakers might have their own interpretation of how to express basic emotions in the L2. Therefore, we could assume that L2 speakers might not be able to produce basic emotions in their L2 as effectively as they do in their L1, although basic emotions should have contained more universal acoustic cues than non-basic emotions, according to Darwin's evolution theory. Therefore, we could further assume that emotion production in an L2 depends very much on the L2 speaker's understanding of the L2, even for some basic emotions, such as 'anger'. Furthermore, the results suggest that production of emotional prosody in one's L1 is the combination of universal acoustic cues and culture-or-language-specific variables.

In the perception of native-produced Chinese emotional prosodies, Chinese native listeners are not able to identify emotions more accurately than naïve Dutch listeners

and advanced Dutch learners of Chinese did. Surprisingly, advanced Dutch learners of Chinese recognize emotional prosody in Chinese significantly better than the natives do themselves. This finding contradicts the conclusion of Graham's (2001) study that the ability to accurately identify emotions being portrayed through vocal cues in a second language may not be acquired by L2 learners without extensive exposure to such emotions in a native context or without special attention to developing these skills in an instructional context. Moreover, advanced Dutch learners of Chinese can identify Chinese emotional prosody significantly (and substantially) better than naïve Dutch listeners. This finding confirms the result of Shoshi and Gagnié's (2010) study that trained second language learners recognize emotional prosody in the target language better than listeners with no experience in the target language.

There may be several possible explanations for the findings that Dutch L2 speakers of Chinese were not able to produce emotional prosody in their L2 (i) as well as Chinese natives and (ii) as successfully as in their L1.

First of all, Ross et al. (1986) have shown that there is less use of short-term changes in F0 to express emotion in tone languages (in which short-term F0 contours are used to carry lexical information) than in Indo-European languages (in which F0 typically plays no lexical role). Thus it seems that in some cases at least, use of a particular acoustic feature in spoken language limits its use for the communication of emotion. Inspired by Ross et al., one might predict that the prosodic space which languages may use is finite. The parameters (or dimensions) of the phonetic space (and the prosodic space within it) can be used to express linguistic as well as paralinguistic contrasts. In other words, if a language uses a prosodic parameter for linguistic purposes, it can no longer use the same parameter for non-/paralinguistic uses – or, in a less extreme version of the theory – cannot use the same parameter as effectively for the expression of paralinguistic or extralinguistic meanings. The prediction follows that speakers of a lexical tone language (such as Mandarin) have less room to express emotion through prosody (specifically through paralinguistic use of speech melody) than speakers of a non-tone language (such as Dutch or English). Apparently, native Chinese speakers can pronounce Mandarin lexical tones correctly without thinking during the production of emotional prosody in their native language, but L2 speakers of Chinese cannot. In this case, L2 speakers of Mandarin are not able to easily separate emotional prosody from lexical tones during their production of Chinese emotional prosody, so that they cannot express it as well as natives. It can also explain why Dutch L2 speakers cannot vocally produce emotions as well as they do in their L1. As a consequence of the prediction, listeners of a tonal language will be less intent on (and in fact less experienced in) decoding the paralinguistic use of prosody than listeners of a non-tonal language. In other words, listeners of a non-tonal language are generally better at recognizing emotional prosody than listeners of a tonal language. This would explain why naïve Dutch listeners can recognize Chinese emotional prosody as well as natives, and why advanced Dutch L2 learners of Chinese can identify the same emotions even better. It is worth rerunning this experiment with different linguistic groups to see if the results are similar, for example, British naïve listeners and British L2 learners of Chinese; or German naïve listeners and German learners of Chinese.

Secondly, Chinese society is quite reserved when it comes to the overt expression of emotion, either in speech or in other modes of communication (Klineberg 1938). Showing emotion in public is interpreted as a sign of weakness in China (Wu & Tseng 1985). If this is indeed the case, then native speakers of Chinese will have had little exposure to clear instances of vocally expressed emotions. This would explain why Dutch L2 speakers could not produce emotional prosody in their L2 as well as natives, which is simply due to the same reason that they lack clear input of exemplars of emotional prosody produced in Chinese.

In order to better understand the production of emotional prosody in the speakers' L2 and L1, I carried out an acoustic analysis, which is presented in the next chapter. Three groups of speakers, i.e., L1 Dutch speakers, L2 Mandarin speakers and L1 Mandarin speakers (the former two are the same individuals), will be studied in this chapter.

# Chapter Six
# Acoustic Analysis

**Abstract**

This chapter presents an acoustic investigation of emotional prosody produced by three types of speakers, i.e., L1 Dutch speakers, L2 Mandarin speakers and L1 Mandarin speakers, the former two of which are comprised of the same individuals.[16] Eight acoustic correlates were examined in this chapter: mean utterance duration (tempo), mean F0, Standard Deviation of F0, slope of the F0, spectral compactness, Standard Deviation of intensity, jitter (a measure of cycle-to-cycle pitch variation) and HNR (Harmonics to Noise Ratio, a measure of breathiness). The acoustic analysis shows that F0 is a crucial factor in the production of emotional prosody, regardless of speaker type; other acoustic variables are emotion specific or speaker-type specific. Moreover, the acoustic analysis indicates that the Dutch L2 speakers of Chinese have developed a hybrid system to vocally express emotional prosody in their L2-Chinese. This (L2) hybrid system approximates to some extent the Chinese native manner of portraying vocal emotion (the way it involves utterance duration, mean F0, slope of the F0, compactness and jitter), but exploits the variability in F0 and intensity that the L2 speakers use to produce emotional prosody in their L1. I have also performed automatic recognition, by Linear Discriminant Analysis (LDA), of the six emotional prosodies portrayed by the three speaker groups after the acoustic analysis. The automatic recognition aims to find out to what extent the acoustic analysis reflects the human perception of the vocal emotions. The results show that the LDA reflects human perception of emotional prosody to some extent; however, human perception is still different from the computer perception.

[16] This chapter is the second part of Y. Zhu (2013). Production of emotional prosody in L2 and in L1 (submitted).

## 6.1 Introduction

In this chapter, I will acoustically analyze selected stimuli from the first two judgment studies (chapter 5), including the six Chinese emotional prosodies expressed by the four native speakers and the four Dutch L2 speakers of Chinese, as well as the six Dutch emotional prosodies produced by the same Dutch L2 speakers of Chinese. The acoustic analysis will answer the following questions:

(1) What acoustic parameters contribute to differentiating between emotional prosodies in general?
(2) What acoustic correlates are used in a language-specific fashion in the production of emotional prosodies
    a. by native Chinese speakers in Chinese
    b. by Dutch L2 speakers of Chinese
    c. by Dutch speakers in the production of Dutch?
(3) Do the Dutch L2 speakers use L1-transfer to vocally produce emotion in their L2 Chinese?
(4) To what extent does automatic recognition reflect the perception of the emotional prosodies by the three groups of human listeners?

Two Mandarin stimuli were excluded (see Table 5.1), since these were not well perceived by the three listener groups in the first judgment study. So the final stimulus sets for the acoustic analysis are equal in size: four Mandarin statements and four Dutch statements were used. Therefore, there are in total 6 vocal emotions × 4 sentences × 4 speakers × 3 speaker types = 288 stimuli available for acoustic analysis. The acoustic analysis was conducted in a comparative way where speaker types (in which language an emotion was expressed), acoustic variables and emotions were presented in the same figure. There is also automatic recognition of the six emotional prosodies portrayed by the three speaker groups after the acoustic analysis. The automatic recognition aims to find out to what extent the acoustic analysis reflects the human perception of the vocal emotions. If the identification rate of the automatic recognition (or the confusion structure) is close to that of the human perception, it would very likely that the acoustic variables the computer used to identify the emotional prosodies are also used by humans.

The first two judgment studies confirmed previous literature that listeners are rather good at inferring affective state and speaker attitude from vocal expression (Frick 1985, Scherer 1986, Standke 1992, Van Bezooijen 1984). Scherer (1996) claimed that listener-judges are able to recognize reliably different emotions on the basis of vocal cues alone, which implies that the vocal expression of emotions is differentially patterned. According to Scherer's (1996) review, previous studies of emotional prosody have examined the following acoustic variables which are strongly involved in vocal emotion signaling:

a) the level (mean F0), range (difference between 95$^{th}$ and 5$^{th}$ percentile), and contour of the fundamental frequency (referred to as F0; it reflects the frequency of the vibration of the vocal folds and is perceived as pitch);
b) the vocal energy (or intensity, perceived as loudness of the voice);

c) the gross distribution of energy in the frequency spectrum (particularly the relative energy in the high vs. the low-frequency region, affecting the perception of voice quality or timbre);
d) the location of the formants (F1, F2…F*n*, related to the perception of articulation); and
e) a variety of temporal phenomena, including tempo and pausing.

Therefore, I am going to look at the following acoustic variables obtained from computer analyses of the speech signals, which will be explained in more detail in the following sections:

(1) tempo (normalized utterance duration);
(2) mean fundamental frequency for the entire utterance, standard deviation of F0 and the difference in mean F0 during the first and last quarter of the utterance duration, which difference is named 'slope' in the present chapter;
(3) the distribution of energy in four contiguous frequency bands from which we will derive a spectral 'compactness' measure;
(4) variation in vocal energy, expressed by the standard deviation of the intensity;
(5) mean jitter;
(6) mean Harmonics to Noise Ratio (HNR or harmonicity).

## 6.2 Acoustic analysis of the selected stimuli

### 6.2.1 Acoustic analysis

#### 6.2.1.1 Utterance duration

I decided to use utterance duration as an approximation to speaking rate, in order to show differences between L1 and L2 speakers vocally expressing emotions in their L1 and L2. Although the stimuli were spoken in two different languages (Mandarin and Dutch) by different speaker types, it is still possible to make a comparison between speaker types across the emotions, as the length, the syllables and the syntactic structure (including pauses) of each stimulus were very well matched between the two different languages (see Tables 5.1 and 5.2). Very often researchers use utterance duration as a first step toward computing tempo measures such as speech rate (syllables/s including pause into the utterance duration) or articulation rate (syllables/s not counting pauses). I preferred to keep the utterance as an integral prosodic unit. Since we are interested in the effects of intended emotion on speaking rate, there is no need to divide the utterance duration by the number of linguistic units contained by it. Instead, it is more convenient to abstract away from the internal linguistic make-up of the various utterances by applying z-normalization within speakers and within lexical sentences, so that only differences between emotions remain as a factor influencing the z-score. This procedure allows us to make direct comparisons of utterance durations between native Mandarin, Dutch L2 Mandarin and native Dutch emotional utterances.

Figure 6.1 shows the mean z-transformed utterance duration of stimuli for the six emotions sorted by the language in which the emotions were expressed. The emotions

are plotted along the horizontal axis. The three speaker groups are represented in different panels. As can be seen from Figure 6.1, both Dutch L2 and native speakers of Chinese used slower speed (i.e. longer utterance duration) to express 'sadness' and 'sarcasm' in Mandarin. However, Dutch L2 speakers of Chinese did not slow their speaking rate to portray 'sadness' in their L1. Moreover, L1 Chinese speakers tended to talk faster when they were angry or happy; but this tendency was only seen with Dutch L2 speakers of Chinese producing 'happiness' in their L1. Overall, the signaling of emotion by variation in utterance duration by the Dutch L2 speakers of Chinese is less outspoken (i.e. smaller differences between the six emotions) than in the L1 of either the same Dutch speakers or in the L1 of the Chinese control group.

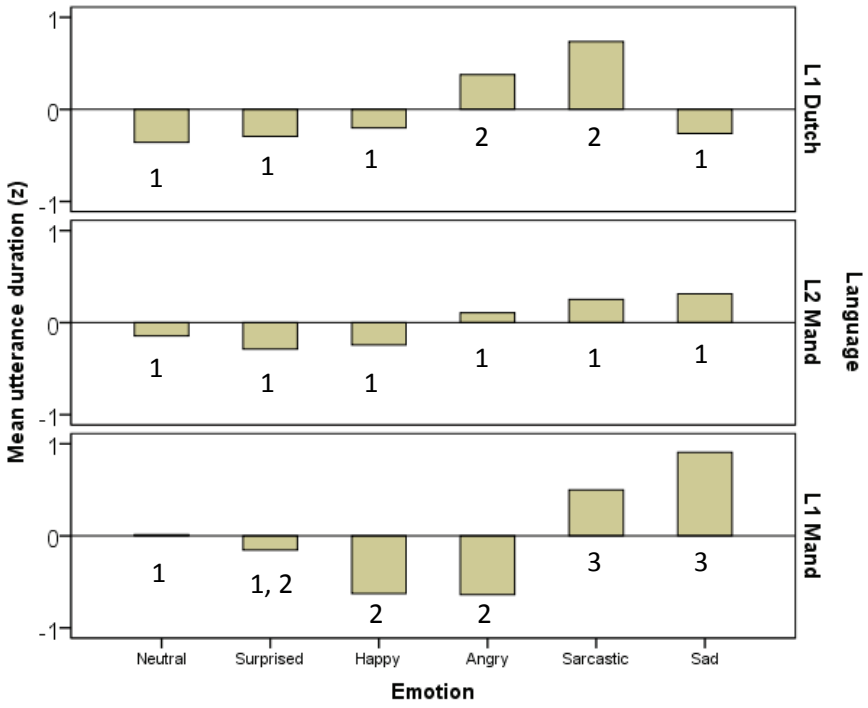



Figure 6.1. *Mean utterance duration (z-normalized within speakers) of stimuli across six emotions, classified by speaker type. 'L1 Mand' = Mandarin spoken by Chinese native speakers; 'L2 Mand' = Mandarin spoken by Dutch L2 speakers; 'L1 Dutch' = Dutch spoken by Dutch L2 speakers of Chinese. L1 Dutch and L2 Mandarin are the same individuals. Emotions within the same panel sharing the same group number do not differ significantly from each other (Bonferroni post-hoc procedure).*

According to a oneway ANOVA and Bonferroni post-hoc procedure (see Appendix 2.1 for detailed results), there is no significant effect with the L2 Mandarin speakers, $F(5, 90) = 1.1$ ($p = .360$, ins.), meaning that no emotions differ from each other in

terms of tempo. The same procedure indicates a significant effect of emotion for the L1 Dutch speakers, $F(5, 90) = 3.8$ ($p = .003$); 'sarcasm' differs from all other emotions except 'anger' but no other contrasts are significant. The effect of emotion is also significant for the L1 Mandarin speakers, $F(5, 90) = 9.8$ ($p < .001$); 'sad' is slower than all other emotions, while 'sarcastic' is additionally slower than 'happy' and 'angry', which do not differ from each other.

#### 6.2.1.2 Fundamental frequency (F0)

The fundamental frequency (F0) of the voice represents the frequency of the vibration of the vocal folds during phonation (Scherer 1991). Three parameters were extracted for each emotional utterance in the database, i.e. the mean F0, standard deviation of F0 and slope of F0. F0 was measured using the autocorrelation method implemented in the Praat speech processing software (Boersma & Weenink 1996). For each speaker appropriate cut-off frequencies were established by trial and error. F0 was measured in hertz (Hz) for 10-ms frames. Resulting pitch tiers were visually inspected and obvious errors were corrected interactively. Mean and standard deviation of F0 were then computed as the arithmetic mean and SD of the (corrected) Hz-values for all voiced analysis frames. The SD of the fundamental frequency (SD_F0) captures the overall variability in fundamental frequency over the course of an utterance. One can imagine that some emotions (e.g. 'surprised') are characterized by large pitch movements – and therefore by a large SD_F0 – while others tend to have a rather flat pitch (such as 'sad') with a low SD_F0. Mean and SD of the F0 are not enough to characterize the overall trend in the pitch curve of an utterance. Therefore I added a third pitch-related parameter in order to specifically capture the rising or falling trend in the F0 over the course of the utterance. The F0-slope was computed by taking the difference between the mean F0 computed (as defined above) for the first quarter of the utterance duration and during the last quarter. The slope thus captures the gross rising or falling nature of the sentence melody over the course of the utterance. If the mean F0 is higher in the final quarter than in the first, the melody is basically a rising pattern with an upward slope (with a positive value, as could be expected in the case of surprise); if the last quarter is lower than the first, the melody is a fall (with a negative, i.e. downward slope, as would be expected in the case of a neutral statement or of a sarcastic utterance).

A problem in the comparison of the three speaker groups is that they are composed of different numbers of male and female speakers. One way to deal with this is to present the results separately for each of the genders. An alternative would be to normalize the F0 measurements on a speaker-individual basis by converting the F0 measurements to z-scores such that each speaker – whether male or female – has a mean F0 of 0 and a standard deviation of 1. For the Dutch speakers the normalization was applied separately for Dutch emotions and for L2 Mandarin emotions (as if the L1 Dutch speakers and the L2 Mandarin speakers were different groups of individuals). The reason to run the normalization separately per language is that the Mandarin materials have different lexical structures (with tones in the case of the Mandarin materials) so that differences in mean pitch or 'slope' would not be meaningful when compared across languages. The same normalization was carried out for the SD_F0 and the 'slope' parameters. The effects for the three variables are shown in Figures 6.2-3-4,

respectively, broken down by speaker type (native Mandarin, native Dutch, Dutch speakers of Mandarin) and intended emotion.

Figure 6.2 presents the z-normalized mean F0 values for the six emotions (along the horizontal axis) produced for the three speaker groups (in separate panels).

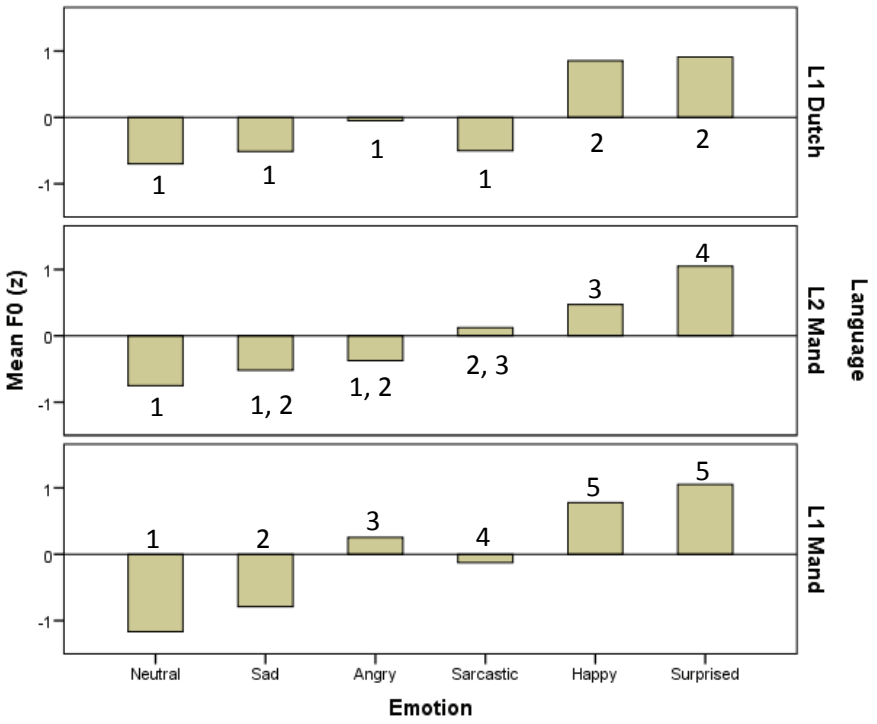



Figure 6.2. *Mean F0 (z-normalized within speakers) for six emotions broken down by speaker group (further see Figure 6.1).*

Figure 6.2 shows that the L1 Mandarin speakers make a very systematic difference in mean F0 between emotions. The six emotions show a monotonically increasing mean F0 when ordered neutral < sad < sarcastic < angry < happy < surprised. The increments in z-scores are roughly equal within any adjacent pair of emotions. The same ordering is found for the L1 Dutch speakers but the increments between adjacent positions are less regular. The effect of emotion on mean F0 is very strong for the L1 Mandarin speakers, $F(5, 90) = 41.7$, $\eta^2 = .95$ ($p < .001$). A Bonferroni post-hoc analysis ($\alpha = .05$) shows that all emotions differ significantly with the exception of 'happy' and 'surprised', which do not differ from each other. The effect of emotion is considerably smaller for the L1 Dutch speakers, $F(5, 90) = 14.5$, $\eta^2 = .70$ ($p < .001$); here 'happy' and 'surprised' do not differ from each other but have higher mean F0 than all other emotions, which do not differ from each other. The effect of emotion is smallest for Dutch speakers of Mandarin, $F(5, 90) = 12.1$, $\eta^2 = .62$ ($p < .001$); here 'surprised' is

higher-pitched than any other emotion, while 'neutral' is lower-pitched than 'sarcastic' and 'happy' (for more details see the subgroup structure indicated numerically in Figure 6.2 and Appendix 2.2).

In terms of mean F0, it would seem then that four emotions do not differ much between Dutch and Mandarin (presumably sharing the universal part of the code). 'Happiness' and 'surprise' are expressed through high pitch in both languages whereas 'neutrality' and 'sadness' are low-pitched. A difference between Mandarin and Dutch is seen in the coding of '(hot) anger' and 'sarcasm'. 'Sarcasm' is low-pitched in Dutch but pitch-neutral in Mandarin. Interestingly, the Dutch learners of Mandarin seem to have picked this language-specific cue, since they have replaced the Dutch low pitch by neutral pitch when they try to be sarcastic in Mandarin. As for 'anger', the L2 Mandarin speakers have opted for an incorrect strategy here: their low-pitched 'anger' in Mandarin deviates from what they do in Dutch but also from what native speakers of Mandarin do.

In very much the same way I analyzed the effects of emotion on the standard deviation of the fundamental frequency, SD_F0. The details are graphically presented in Figure 6.3 (for the subgroups, see Appendix 2.3).

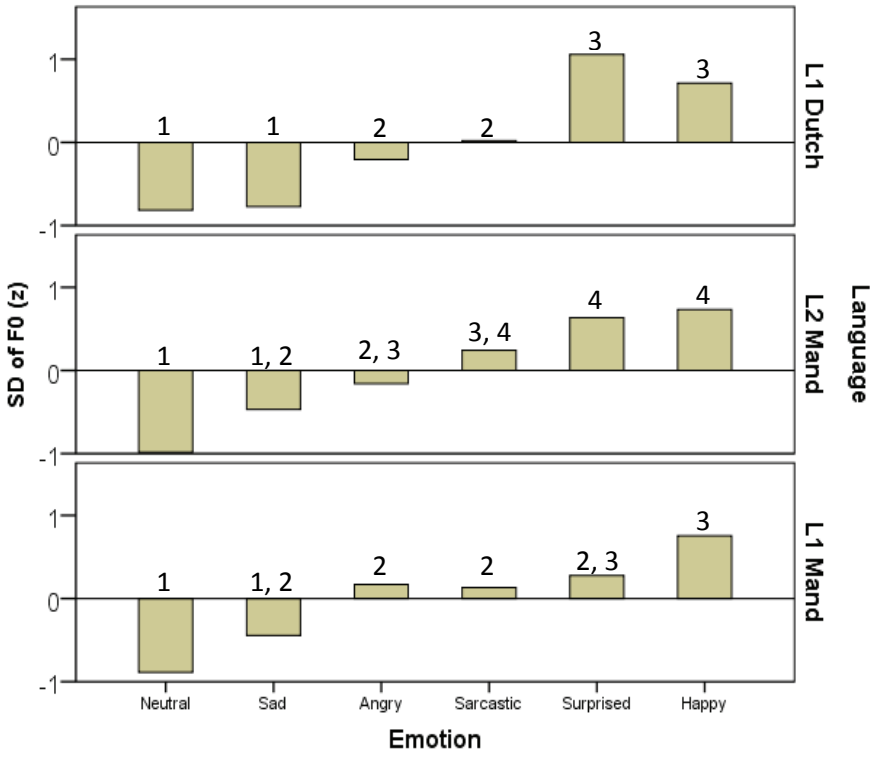



Figure 6.3. *Standard deviation of F0 (z-normalized within speakers) for six emotions broken down by speaker group (further see Figure 6.1).*

Emotion had a highly significant effect on the SD_F0 for the Mandarin L1 speakers, $F(5, 90) = 7.4$ ($p < .001$). 'Neutral' obtained the lowest SD_F0 value and significantly differed from all other emotions except 'sad'. 'Happy' was characterized by the largest SD_F0 and differed not only from 'neutral') but also from 'sad'. The effects are stronger for the Dutch native speakers, $F(5, 90) = 18.7$ ($p < .001$). The six emotions are characterized by SD_F0 in almost the same order as with the Mandarin speakers (neutral < sad < angry < sarcastic < happy < surprised) but the differences between (groups) of emotions are stronger: 'neutral' and 'sad' have low SD_F0 and differ from all other emotions, 'angry' and 'sarcastic' are in a middle group and differ from all others, and 'happy' and 'surprised' have the highest SD_F0 values, differing from all others. The effect is intermediate for the Dutch L2 speakers of Mandarin, $F(5, 90) = 11.3$ ($p < .001$). The order of the emotions is virtually the same as when these speakers produce them in their L1, with an insignificant reversal of 'surprised' and 'happy' in the top SD_F0 group only. However, there is more overlap between the emotion groupings.

Figure 6.4. presents the effects of emotion on the gross slope of the fundamental frquency contour over the course of the utterance (slope_F0) for the three speaker groups.

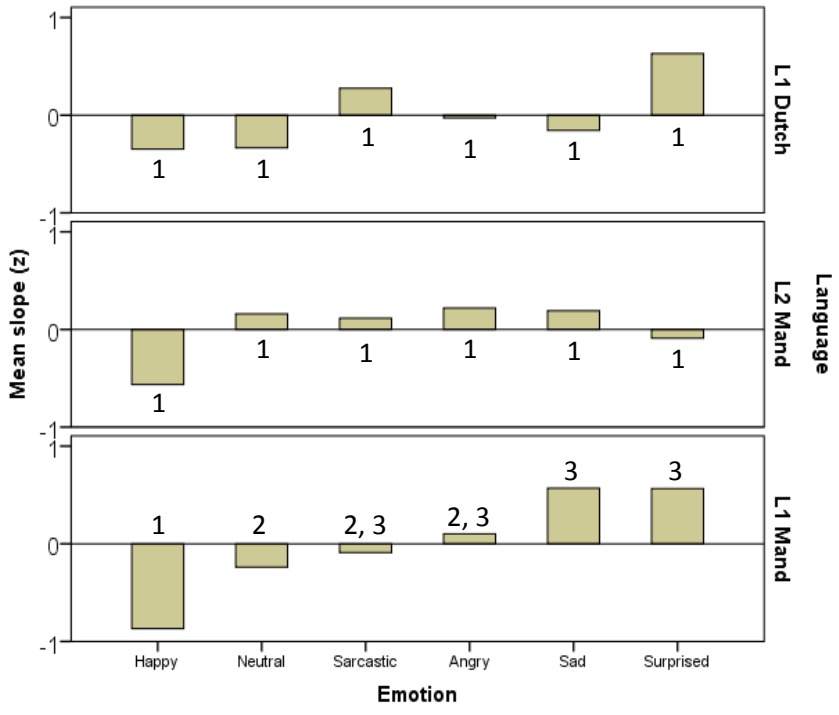



Figure 6.4. *Mean F0 slope (z-normalized within speakers) for six emotions broken down by speaker group (further see Figure 6.1).*

Figure 6.4 indicates that the slope measurement is sensitive to emotion only for the native Chinese speakers, $F(5, 85) = 7.1$ ($p < .001$).[17] There are no significantly different groups among the emotions with L1 Dutch speakers, $F(5, 84) = 2.3$ ($p = .035$) and L2 Chinese speakers (who are the same individuals), $F(5, 76) = 1.4$ ($p = .233$, ins.) (see Appendix 2.4 for detailed results). According to the Bonferroni post-hoc procedure 'surprised' is characterized by a rising pitch, and differs significantly from 'neutral' and 'happy', both of which have falling pitch (but 'happy' significantly more so than 'neutral'. This finding confirms previous studies (e.g. Yip 2006) that many tonal languages use rising intonation to express surprise.

#### 6.2.1.3 Compactness

In order to compute a measure capturing the compactness of the spectral distribution, mean intensity was measured (in dB) in each of four contiguous frequency bands: b1 (0-500 Hz), b2 (500-1000 Hz), b3 (1000-2000 Hz) and b4 (2000-4000 Hz). Following Van Santen et al. (2009) we defined compactness as the difference between (b2 + b3) minus (b1 + b4). When energy is concentrated in the middle of the spectrum, the compactness measure is relatively high and positive, when energy is rather more distributed over low and high frequencies (leaving less energy in the middle portion of the spectrum), the compactness measure is close to zero or even assumes negative values. This compactness measure showed very clear contrasts between at least 'happiness' and 'anger' in Van Santen et al.'s study. In order to be able to make an unbiased comparison across the three speaker groups (with different numbers of male and female speakers) we z-normalized the compactness measure within languages and within individual speakers. The normalized compactness values for the present experiment as shown in Figure 6.5, sorted by emotion and by speaker type.

[17] In a number of cases no mean F0 could be established for either the first or the last quarter of the utterance (or even both). In such cases no slope measure was computed, leaving a smaller number of valid cases for the ANOVA.

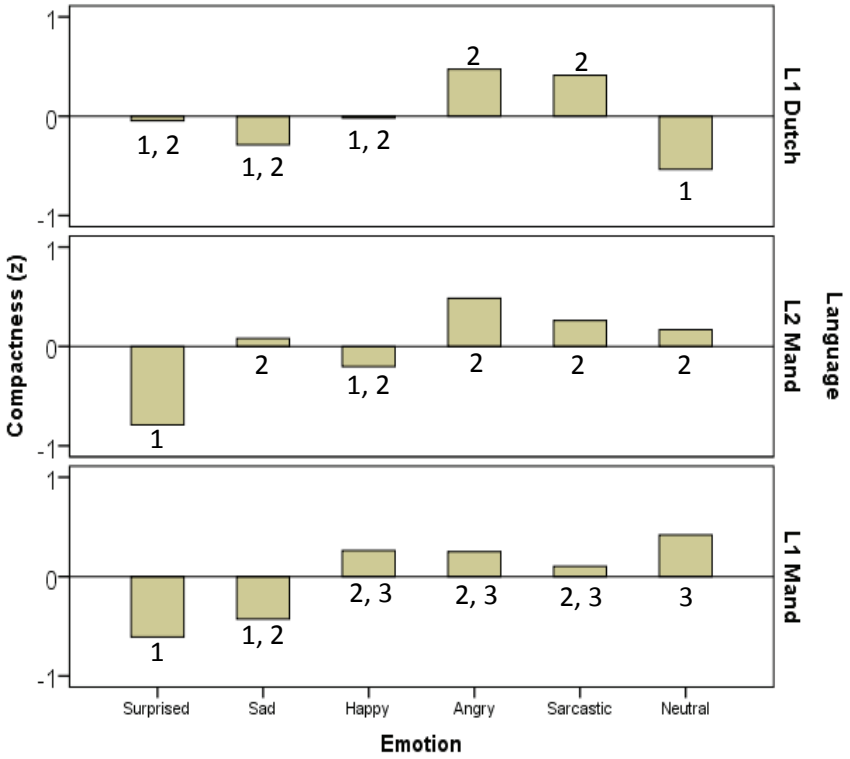



Figure 6.5. *Mean compactness (z-normalized within speakers) across the six emotions, classified by speaker group (further see Figure 6.1).*

Figure 6.5 shows that the compactness measure is sensitive to emotion in all the speaker groups. I applied the same statistical method as before, i.e. a one-way ANOVA to establish the overall effect of the factor intended emotion followed by Bonferroni post-hoc tests ($\alpha = .05$) to determine the statistical difference between each of the six emotions. For the L1 Dutch speakers, the effect of emotion is significant, $F(5, 90) = 2.8$ ($p = .023$); two emotions obtain a positive z-value, i.e. 'anger' and 'sarcasm'. These two emotions differ significantly only from 'neutrality' (which obtains the lowest negative z-value. No other differences are significant. A slightly stronger effect, $F(5, 90) = 3.8$ ($p = .004$) of emotion is seen with L2 Mandarin speakers. Here only surprise (with a negative z-value) differs from all other emotions, which do not differ from each other. Finally, the effect of emotion is also significant for the L1 Mandarin speakers, $F(5, 90) = 4.4$ ($p = .001$); here 'surprised' differs from all emotions except 'sad' while 'neutral' differs from 'sad' and 'surprised'. In all there is substantial overlap among the emotions, as can be seen in Figure 6.5 (see Appendix 2.5 for detailed results). This implies that the single measure of compactness as proposed by Van Santen et al. (2009) does not afford an effective division of emotions in our recordings.

### 6.2.1.4 Intensity

Scherer (1991) claimed that 'intensity is a difficult variable to measure since it depends highly on the distance and direction of the speaker's mouth to the microphone, the gain setting of the tape recorder, the equipment used, etc.' In order to circumvent this problem I did not measure (mean) intensity per utterance but concentrated on the variation in intensity around the mean per utterance. This would then provide us with a handle on the dynamic nature of the speaker's voice. When there is little variation in intensity over the course of the utterance the speaker makes little difference between loud and weak syllables. Large variability would characterize an utterance with large differences between loud and weak syllables (or larger units). The variability measure I adopted is the standard deviation of the intensity in the utterance. The results are presented in Figure 6.6.

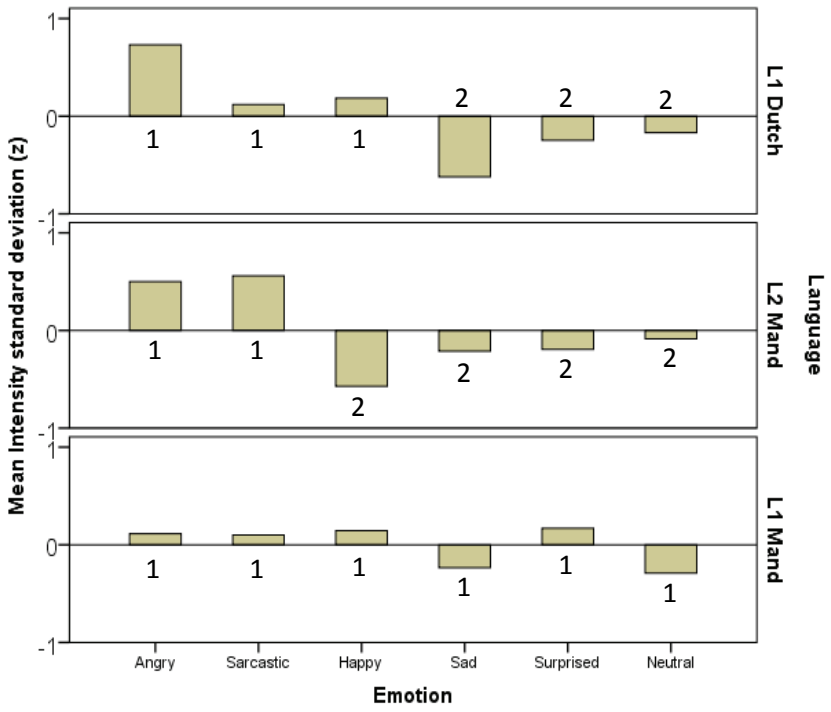



Figure 6.6. *Mean standard deviation of intensity (z-normalized within speakers) across the six emotions sorted by speaker group (further see Figure 6.1).*

As can be seen from Figure 6.6 (see Appendix 2.6 for details of the post-hoc analysis), L1 Mandarin speakers tended to portray all the emotions with little gradation in intensity, $F(5, 90) < 1$ and none of the emotions differs from any of the others. Figure

6.6 also indicates that the effect of emotion is significant for L2 Mandarin speakers, $F(5, 90) = 3.7$ ($p = .004$), and for the same speakers talking native Dutch, $F(5, 90) = 4.1$ ($p = .002$). In their L2 Mandarin 'happy' is significantly flatter than both 'angry' and 'sarcastic', which do not differ from each other. The liveliness of 'sarcasm' gets lost when the same speakers speak native Dutch: here only 'angry' is more lively (less flat) than any other emotion. So, it would appear that the Dutch L2 Mandarin speakers speak relatively evenly when they express 'happiness', 'sadness', 'surprise' and 'neutrality' in their L2 as they do in their L1 except 'happiness'.

#### 6.2.1.5 Jitter and Harmonics-to-Noise Ratio

Jitter and Harmonics-to-Noise Ratio (HNR) are another two frequently studied parameters which are believed to contribute to perception and production of emotional prosody. Jitter, also known as pitch perturbation, refers to the minute involuntary variations in the frequency of adjacent vibratory cycles of the vocal folds. Excessive jitter makes a voice sound rough and unstable. This measurement can tell us how creaky or rough an emotional prosody is, especially when a speaker has a weeping voice while producing the emotion (e.g. sadness). I used the ppq5 jitter measure which is one of the jitter measures implemented in the Praat speech processing software. This measure computes the pitch perturbation coefficient as the mean of the differences between successive periods in a five-period window, divided by the mean period in the same window (for details see Davis 1976, Kraayeveld 1997, Pinto & Titze 1990). This yields a coefficient between 0 (absence of any jitter) and 4 (extreme, pathological, roughness).

The Harmonics-to-Noise Ratio (HNR) is used to measure the hoarseness of a voice. According to Speech Therapy Information and Resources (2008), 'the aperiodic waves are random noise introduced into the vocal signal owing to irregular, asymmetric or incomplete adduction (closing) of the vocal folds. Noise impairs the clarity of the voice and too much noise is perceived as breathiness or even hoarseness.' Praat measures the intensity of the harmonics in the (quasi-) periodic parts of the speech wave and of the parts of the spectrum between the harmonics. The intensity difference between the harmonics and the noise between the harmonics is expressed as the Harmonics-to-Noise (HNR) ratio (in dB). A clear voice is the characterized by a large positive HNR value (i.e. there is hardly any noise between the harmonics); a breathy, and especially a hoarse voice has a low or even negative HNR (the latter indicating that the noise between the harmonics is even louder than the harmonics themselves). Therefore, it is worth looking at these two parameters in the acoustic analysis of the emotional prosodies. The z-normalized mean jitter and HNR values are presented in Figure 6.7 and 6.8, respectively (for details of the post-hoc analysis, see Appendices 2.7 and 2.8).

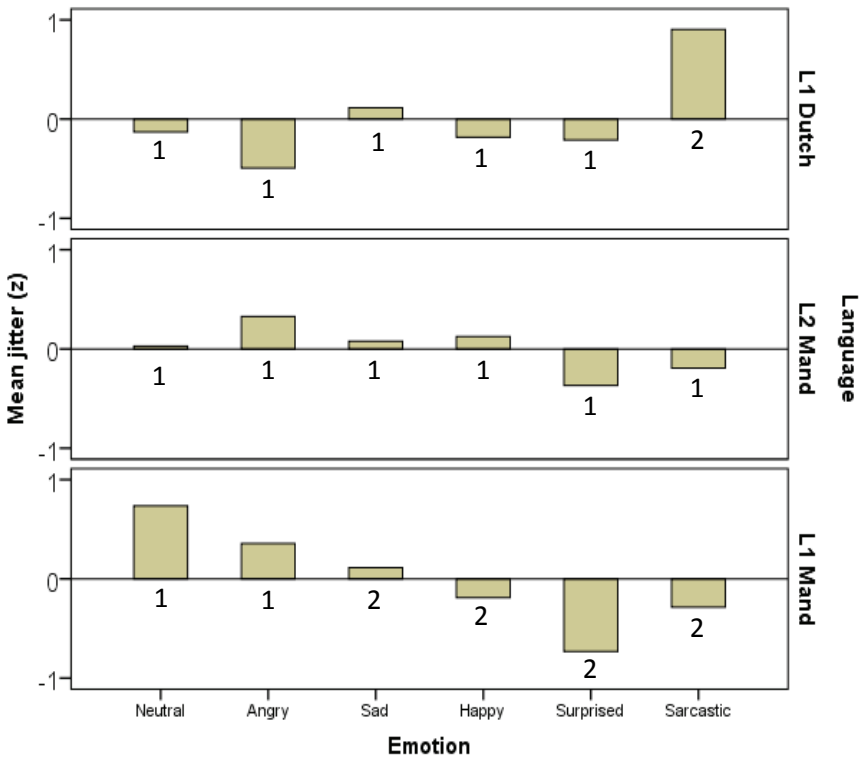



Figure 6.7. *Mean jitter (z-normalized within speakers) across the six emotions sorted by speaker group (further see Figure 6.1).*

As can be seen from Figure 6.7, there are two groups which significantly differ from each other among the emotions with L1 Dutch and L1 Mandarin speakers respectively. Specifically, 'sarcasm' is separated from other emotions (which do not differ from each other at $\alpha = .05$) by a larger jitter measure with L1 Dutch speakers, $F(5, 90) = 4.57$ ($p < .001$). Emotion also has a significant effect on jitter in the speech of L1 Mandarin speakers, $F(5, 90) = 6.02$ ($p < .001$). Here 'surprise' has lower jitter than all other emotions while 'neutral' has more jitter than all other emotions (which do not differ between them). This finding indicates that L1 Dutch and L1 Mandarin speakers portrayed the emotions in a very different way in terms of vocal stability. However, the jitter measure is not sensitive to emotion for L2 Mandarin speakers where there is no emotion that differs significantly from any of the others, $F(5, 90) = 1{,}00$ ($p = .422$, ins.).

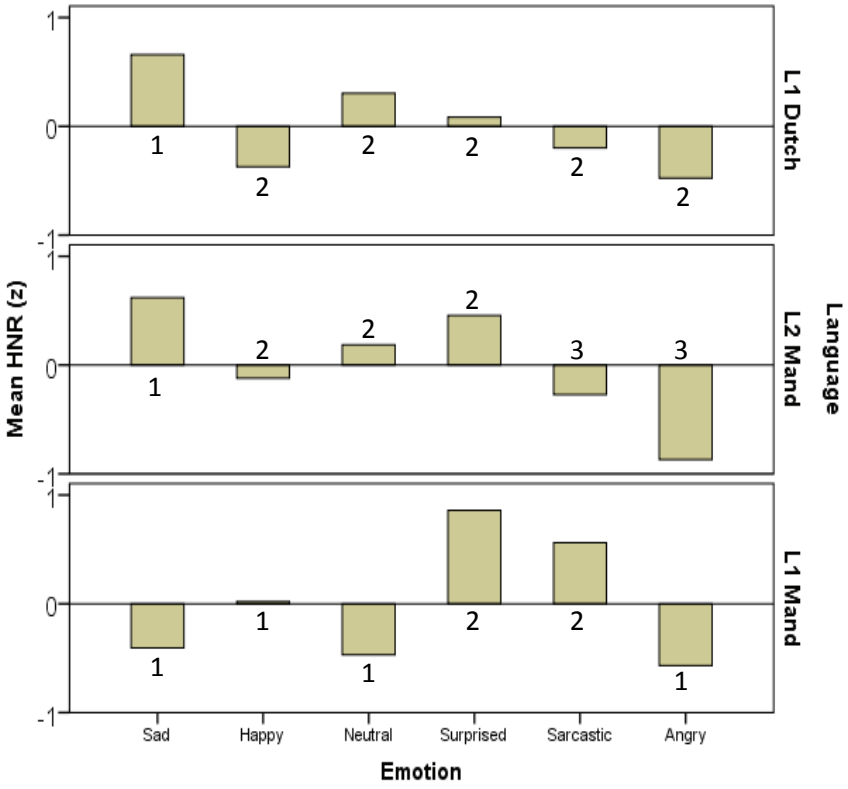



Figure 6.8. *Mean HNR (z-normalized within speakers) across the six emotions sorted by speaker group (further see Figure 6.1).*

Figure 6.8 shows that the HNR measure is sensitive to emotion for all three speaker groups. The results for the L1 Dutch speakers show that 'sadness' differs from the other five emotions by having a better (i.e. less noisy) HNR; 'angry', 'happy' and 'sarcastic', which have relatively poor HNR, differ from all other emotions but not from each other, $F(5, 90) = 3.50$ ($p = .006$). For the L1 Mandarin speakers 'angry', 'neutral' and 'sad', which do not differ from each other, differ from all other emotions by their lower harmonicity; 'surprised' does not differ from 'sarcastic' but differs from all other emotions by its higher harmonicity, $F(5, 90) = 7.80$ ($p < .001$). With the L2 Mandarin speakers 'angry' does not differ from 'sarcastic' but differs from all other emotions by its lower HNR-value. 'Sad' has the highest HNR-value and differs from both 'angry' and 'sarcastic' but not from any other emotions, $F(5, 90) = 6.17$ ($p < .001$).

### 6.2.2 Automatic computer recognition of the six emotional prosodies

In this section, I will report on an attempt at automatic recognition of the six emotional prosodies portrayed by the three speaker groups: L1 Dutch speakers, L2 Mandarin speakers and L1 Mandarin speakers (where the former two groups are in fact the same individuals). The automatic recognition made use of all the acoustic variables discussed and analyzed above, including (a) tempo; (b) mean, standard deviation and 'slope' of the fundamental frequency; (c) spectral compactness; (d) vocal energy (standard deviation of intensity); (e) jitter (ppq5); (f) HNR. These eight acoustic measures were used as predictors in a Linear Discriminant Analysis (LDA, for background of the technique see e.g. Klecka 1980) classifying a total of 288 tokens (utterances) into the six emotional categories separately for each of the three speaker groups (96 tokens per speaker group). The analysis was run in stepwise mode (with default parameter settings for inclusion and exclusion of predictors), in order to force the algorithm to come up with an optimal (most economical) solution of the classification task. In this application of the LDA the algorithm was trained and tested on the same data; no attempt was made to cross-validate the solution. The results of the LDA are presented in the form of a confusion matrix, with the intended emotions as the stimulus variable (in the rows) and the emotions as predicted (classified) by the LDA as the response variable (in the columns). Correctly classified emotions are on the main diagonal; confusions are in the off-diagonal cells. I will present the confusion matrices separately for each of the three speaker groups.

Table 6.1 shows the perception results of the LDA for the three speaker types. The overall mean recognition rate for the L1 Dutch speakers is 49%, for the L2 Mandarin speakers (who were the same individuals as the L1 Dutch speakers) it is 34% and for the native Mandarin speakers it is 66%. The comparable recognition rates of human listeners in the present study are: 57% (L1 Dutch speakers), 39% (L2 Mandarin speakers) and 48% (L1 Mandarin speakers), regardless of listener type. The correct identification rates by LDA (50% correct) overall and by human listeners (48% correct) for the three speaker groups are similar. These correct identification scores, whether by machine or by human listeners, are about three times better than chance (1/6 = 17% correct).

We may also try to determine the extent to which the confusion structure in the computer identification of the emotions reflects that of the human listeners. The correlation coefficient between the confusions (percentages in the off-diagonal cells only) obtained by the LDA and those by the human listeners (same group as the speakers) was small but significant, $r = .36$, $n = 30$, $p < .05$ (one-tailed) for the L1 Dutch speakers and $r = .33$, $n = 30$, $p < .05$ (one-tailed) for the L2 Mandarin speakers. However, there was no significant correlation between the confusions by LDA and by the human listeners for the L1 Mandarin speakers. These correlation results indicate that the perception of emotional prosody by computer is rather different from that by human listeners in general. Although the LDA can identify human-produced emotional prosodies to some extent, the acoustic correlates that the LDA singles out to classify the vocal emotions are not necessarily those that are used by human listeners.

Table 6.1. *Automatic recognition by LDA of emotional prosody produced by three speaker groups: Confusion matrix of intended and perceived emotions portrayed by L1 Dutch speakers (upper panel), L2 Mandarin speakers (middle panel) and L1 Mandarin speakers (lower panel). Correct responses are located on the main diagonal (shaded). The L1 Dutch and L2 Mandarin speakers are the same individuals. The right-most columns list the mean percentage of correct identifications across all emotions by LDA and by humans across all listener groups (and in parentheses for listeners matching the speaker type).*

| Human intended | Computer Perceived Emotion Encoded by L1 Dutch speakers | | | | | | Mean correct by | |
|---|---|---|---|---|---|---|---|---|
| | Ang | Hap | Neu | Sar | Sad | Spr | LDA | human |
| Angry | **37.5** | 18.8 | 25.0 | 6.3 | 12.5 | 0 | 49 | 57 |
| Happy | 0 | **68.8** | 0 | 0 | 6.3 | 25.0 | | |
| Neutral | 0 | 6.3 | **50.0** | 25.0 | 18.8 | 0 | | |
| Sarcastic | 0 | 18.8 | 18.8 | **56.3** | 6.3 | 0 | | |
| Sad | 25.0 | 6.3 | 56.3 | 6.3 | **6.3** | 0 | | |
| Surprised | 6.3 | 18.8 | 0 | 0 | 0 | **75.0** | | |
| | Computer Perceived Emotion Encoded by L2 Mandarin speakers | | | | | | Mean correct by | |
| | Ang | Hap | Neu | Sar | Sad | Spr | LDA | human |
| Angry | **62.5** | 12.5 | 12.5 | 0 | 12.5 | 0 | 34 | 39 (41) |
| Happy | 25.0 | **12.5** | 0 | 18.8 | 12.5 | 31.3 | | |
| Neutral | 25.0 | 0 | **31.3** | 6.3 | 37.5 | 0 | | |
| Sarcastic | 25.0 | 25.0 | 12.5 | **6.3** | 6.3 | 25.0 | | |
| Sad | 0 | 12.5 | 31.3 | 12.5 | **43.8** | 0 | | |
| Surprised | 0 | 12.5 | 6.3 | 12.5 | 12.5 | **56.3** | | |
| | Computer Perceived Emotion Encoded by L1 Mandarin speakers | | | | | | Mean correct by | |
| | Ang | Hap | Neu | Sar | Sad | Spr | LDA | human |
| Angry | **50.0** | 9.1 | 9.1 | 9.1 | 9.1 | 13.6 | 66 | 48 (46) |
| Happy | 18.2 | **59.1** | 0 | 9.1 | 0 | 13.6 | | |
| Neutral | 0 | 0 | **75.0** | 10.0 | 15.0 | 0 | | |
| Sarcastic | 8.3 | 4.2 | 0 | **75.0** | 8.3 | 4.2 | | |
| Sad | 0 | 0 | 20.0 | 10.0 | **70.0** | 0 | | |
| Surprised | 4.5 | 13.6 | 0 | 4.5 | 4.5 | **72.7** | | |

Furthermore, the Stepwise LDA shows that there are three significant parameters that the algorithm used to discriminate the emotions produced by L1 Dutch speakers: utterance duration, mean F0 and standard deviation of F0. And there are only two parameters that significantly contributed to the automatic recognition of emotional prosody portrayed by L2 Mandarin speakers, viz. mean F0 and HNR. Finally, there are five parameters that the LDA used to discriminate the emotions produced by L1

Mandarin speakers: utterance duration, mean F0, HNR, compactness and F0 slope. Overall speaking, it means that utterance duration, fundamental frequency, HNR, compactness and slope are the main parameters that contribute to the automatic recognition. Possibly, these are also the parameters that human listeners use to perceive emotional prosody, but this is not clearly indicated by the correlation results. However, parameters like jitter and intensity are not the main factors which influence the automatic recognition. I argue that human listeners may use the eight acoustic correlates studied above as cues in perception of emotional prosody in reality, but they may also use some other variables which are not clear at this stage and which are missed in the acoustical analysis.

### 6.3 Conclusions

In the introduction to this chapter I asked four questions, which I will now repeat for convenience sake, and try to answer on the basis of the results obtained from the above analysis.

(1) What acoustic parameters contribute to differentiating between emotional prosodies in general?

In the acoustic analysis I examined the value of eight parameters as correlates of the six emotions studied. The eight parameters were the same for each of the three groups of speakers, i.e. Mandarin L1, Mandarin L2 and Dutch L1 (the latter two were the same individuals). The acoustic analysis shows that fundamental frequency, including mean F0, SD_F0 and slope of the F0, is an influential variable in the production of vocal emotions by the three groups of speakers. This finding confirms the study of Scherer (1996), who claimed that F0 plays a crucial role in the production of emotional prosody. The results also show that jitter and standard deviation of the intensity did not contribute much to differentiating between emotions in the present study. Never were more than two subgroups of the emotional prosodies differentiated for any of the three speaker groups.

The acoustic analysis indicates that F0 plays an important role in the production of emotional prosody generally. Basic emotions such as 'happy' and 'angry' can be clearly discriminated from each other by mean F0 and SD_F0, regardless the speaker type. 'Happy' is characterized by high values for mean and SD of F0 (z-values close to 1) while 'angry' has z-values close to 0. Interestingly, 'neutral' is also universally differentiated from 'happy' and 'angry', viz. by low values for mean and SD of F0 (values close to –1). However, more controlled emotions, e.g. 'surprised' and 'sarcastic', are not well classified by any of the eight parameters examined above. Since 'surprised' includes both positive and negative surprise, the human listeners sometimes misinterpreted this emotion as 'happy' or 'angry', respectively. It indicates that other factors (e.g. personal interpretation of the emotional label) can also influence the perception of vocal emotion.

(2) What acoustic correlates are used in the production of emotional prosodies
   a. by native Chinese speakers in Chinese,
   b. by Dutch L2 speakers of Chinese,
   c. by Dutch speakers in the production of Dutch?

The acoustic analysis shows that 'tempo' and 'compactness' were only sensitive to Mandarin L1 speakers, for whom three subgroups of the emotional prosodies were found. Slope of the F0 indicates that Chinese uses rising intonation to express surprise, which confirms the previous studies, claiming that many tonal languages use rising intonation to express surprise (Yip 2006). Moreover, HNR can clearly distinguish 'sad' from 'neutral' with Mandarin L2 and Dutch L1, who were actually the same individuals, but not in the case of L1 Mandarin speakers.

In summary, fundamental frequency is a very influential variable in the production and perception of vocal emotion in general. Other parameters studied in this chapter also contribute to differentiating between emotional prosodies, but they are more emotion-specific or speaker-type specific. There may be other factors which are also used in the production of vocal emotion in reality but were missed in this chapter. However, production and perception of vocal emotion by humans is a much more complex and integrated procedure. It involves not only acoustic correlations but also other factors, such as, sex, language or personal interpretation of the emotional label.

(3) Do the Dutch L2 speakers use L1-transfer to vocally produce emotion in their L2 Chinese?

The acoustic analysis indicates that Dutch L2 speakers use some acoustic parameters in the production of emotional prosodies in the L2 (Chinese) the same way they do in their L1 (Dutch), e.g. SD_F0 and SD_Int. Therefore, we may conclude that L1-transfer is a strategy for L2 speakers to vocally produce emotions in the L2. However, this strategy may not work for all the emotions, e.g. not for 'surprise' and 'sarcasm'. Moreover, the acoustic correlates the L2 speakers used for portraying vocal emotions in Chinese are not very similar to those used by L1 Chinese speakers. However, L2 speakers of Chinese did not completely adopt their Dutch approach to produce emotional prosody in Chinese. Neither did they fully use Chinese native manner to vocally express emotions in Chinese. Therefore, it seems that the advanced L2 speakers of Chinese have developed a hybrid system of producing emotional prosody in the L2. This (L2) hybrid system approximates to some extent the Chinese native manner of portraying vocal emotion (the way it involves utterance duration, mean F0, slope of the F0, compactness and jitter), but exploits the variability in F0 and intensity that the L2 speakers use to produce emotional prosody in their L1. Emotional prosodies produced in this in-between manner were identified above chance level by both the native and non-native listeners in the present study. However, these emotional prosodies are less recognizable overall (41% correct within-group identification) than those produced in the Chinese native manner (46% correct). This would indicate that the expression of emotion through prosody is limited in an interlanguage. We may speculate that production of emotional prosody in general is universal to some extent, but production

of vocal emotion in L2 is more likely speaker-specific, with greater dominance of the target L2 system as the learner is more advanced.

(4) To what extent does automatic recognition reflect the perception of the emotional prosodies by the three groups of human listeners?

The results of LDA show that the automatic recognition in the present study can identify human-produced emotional prosody well above chance level (50% overall correct). There was significant correlation between confusions obtained by the automatic recognition and by the human listeners in the present study. Moreover, the overall recognition rate of LDA is slightly better than that of the human perception. This indicates that automatic recognition can reflect human perception of emotional prosody to some extent; however, the human perception is still different from the computer perception. There are still acoustic correlates which used by the algorithm to discriminate between emotions but not used by L1 and L2 listeners in reality. In addition, the Stepwise LDA shows that there are four parameters which significantly contribute to the production and perception of emotional prosody: utterance duration, fundamental frequency, compactness and HNR. It is traditionally argued that intensity and jitter are also important factors (e.g. Biersack & Kempe 2005, Scherer 1996), but these two variables did not influence the automatic recognition very much in the present study. However, I suspect that these two variables may be used in the human perception of emotional prosody in reality too. There may also be some other acoustic parameters contributing to the production and the perception of emotional prosody in general, which have been missed in this dissertation. Further studies can acoustically continue investigating production of emotional prosody in general and production of vocal emotion in speaker's non-native language.

# Chapter Seven
# Perception of Emotional Prosody in a Listener's L1 and in an Unknown Language

**Abstract**

This chapter investigates the perception of emotional prosody by native and novice listeners in a reciprocal way. Twenty Chinese and 20 Dutch native listeners without any knowledge of Dutch and Chinese, respectively, identified emotional prosodies in these two languages. The results show that novice Dutch listeners could recognize emotional prosody in the unknown language (Chinese) as well as natives; and they performed significantly better in identifying emotional prosody expressed in their native language (Dutch). Chinese novice listeners, on the other hand, were able to recognize emotional prosody in their L1 only reasonably well but failed to identify vocal emotion in the unknown language (Dutch) above chance level. This finding confirms the existence of the in-group advantage found by other researchers, claiming that listeners generally better recognize emotional prosody produced in their L1 than in an unknown language. Moreover, the results suggest that cross-cultural perception of vocal emotion is not symmetrical, meaning that some cultural group might be generally more sensitive than some other cultural group in the perception of emotional prosody. This finding lends support to the functional view that predicts that listeners of a tonal language should be generally less intent on the perception of vocal emotion than listeners of a non-tonal language. If this view is accepted, the asymmetry in emotion perception may not only be explained from a difference in culture but also as the result of a difference in linguistic structure.

## 7.1 Introduction

Studies on perception of vocal emotions cross-culturally have been widely carried out since Darwin claimed that affective expressions, including those produced via the vocal channel, are veridical (Darwin 1872). Earlier findings obtained in cross-cultural and/or cross-linguistic studies have borne out that the perception of emotion is partly universal and partly language/culture-specific. Perception of some vocal emotions, for instance, 'anger', 'sadness' or 'neutrality', depends on general biological and physiological mechanisms shared by all humans, meaning that listeners will be able to recognize these emotions even if they are expressed in an unknown language. However, some emotions, for example, 'sarcasm', 'disgust' or 'shame', may well be expressed in different ways depending on the native language and culture of the speaker, and may therefore not be successfully identified by listeners from a different linguistic or cultural background. For instance, Albas et al. (1976) asked male Caucasian speakers of English and Amerindian Cree speakers to express the basic emotions of happiness, sadness, anger, and love, in their respective native languages using any words that came to their minds. These speech samples were electronically low-pass filtered and presented to Caucasian and Cree listeners. The results showed that the emotions were recognized above chance level in all cases but better in the listener's native language than in the unknown language. Therefore, perception of emotion is partly universal and partly culture or language specific. The authors suggest that language and culture are crucial factors in the transmission of emotion, even on the nonverbal level, but admit that the data are difficult to interpret because the content of the speech materials used for encoding was not controlled. Van Bezooijen (1984) studied ten emotional prosodies: neutral, disgust, surprise, shame, interest, joy, fear, contempt, sad, and angry. Her study aimed to find out how (Taiwanese) Chinese and Japanese listeners without any knowledge of Dutch, perceived the Dutch emotional prosodies. All three listener groups recognized the Dutch emotional prosodies well above chance level, with scores of 66, 37 and 33% correct for Dutch, Taiwanese listeners and Japanese, respectively. The Asian listeners' identifications correlated at $r = .6$ with the Dutch identification percentages but correlated somewhat more strongly between Japanese and Taiwanese ($r = .7$). The native and non-native identifications were relatively close together for sadness, fear, surprise and anger (< 30 percentage points difference) whilst other Dutch emotions were identified quite poorly: e.g. joy and shame (both 22% correct against 76 and 61% correct for the native listeners). We assume that the communication of the first group of vocal emotions relies very much on a universal code whereas the latter two depend largely on language-specific cues. Moreover, Scherer et al. (2001) report results from a study conducted in nine countries in Europe, the United States, and Asia on vocal emotion portrayals of anger, sadness, fear, joy, and neutral voice as produced by professional German actors. Data show an overall accuracy of 66% across all emotions and countries. Although accuracy was substantially better than chance, there were sizable differences ranging from 74% in Germany to 52% in Indonesia. Generally, accuracy decreased with increasing language dissimilarity from German in spite of the use of language-free speech samples. It is concluded that culture- and language-specific paralinguistic patterns may influence the decoding process. The results of these studies also showed a tendency for vocal emotion to be generally better recognized within the same cultural group. This tendency is called *ethnic bias* by some theorists (e.g. Kilbride & Yarczower 1983, Markham & Wang 1996), claiming that it is possible that recognition

accuracy is higher when emotions are both expressed and perceived by members of the same cultural group. A preferred term for the same tendency is called 'in-group advantage' by other researchers, further claiming that listeners generally better recognize emotional prosody produced in their L1 than in an unknown language. For example, Elfenbein and Ambaby (2002) conducted a meta-analysis which examined emotion recognition within and across cultures based on literature search. They concluded that emotions were universally recognized at better-than-chance levels. Accuracy was higher when emotions were both expressed and recognized by members of the same national, ethnic, or regional group, suggesting an in-group advantage. They also found that recognition of emotion is partly universal and partly cultural-specific. More recently, Thompson and Balkwill (2006) conducted an experiment in which twenty English-speaking listeners judged the emotive intent of utterances spoken by male and female speakers of English, German, Chinese, Japanese, and Tagalog. Identification accuracy was above chance for all emotions expressed in all languages. Across languages, sadness and anger were more accurately recognized than joy and fear. The (English) listeners showed an in-group advantage for decoding emotional prosody, with highest recognition rates for English utterances and lowest rates for Japanese and Chinese utterances. It would also indicate that, again, emotional prosody is decoded by a combination of universal and culture-specific cues. Pell et al. (2009) carried out a similar study in which they compared how monolingual speakers of Argentine Spanish recognize basic emotions from pseudo-utterances ('nonsense speech') produced in their native language and in three foreign languages (English, German, and Arabic). Results indicated that vocal expressions of basic emotions could be decoded in each language condition at accuracy levels exceeding chance, although Spanish listeners performed significantly better overall in their native language (in-group advantage). On the basis of their findings the authors argued that the ability to understand vocally expressed emotions in speech is partly independent of linguistic ability and involves universal principles, although this ability is additionally shaped by linguistic and cultural variables.[18]

Previous studies typically investigated perception of vocal emotion cross-culturally one-way, i.e., vocal emotion encoded in language A was perceived by native listeners and other culture groups B, C, D, etc. (e.g. Scherer et al. 2001). Or, in some cases, culture group A perceived emotional prosody expressed in its L1 and in several other (and often unknown) languages (e.g. Thompson & Balkwill 2006). However, studies which investigated perception of vocal emotion by different cultural groups in a reciprocal manner are rare. In the reciprocal approach both culture groups A and B perceive emotional prosody not only expressed in their own, native language (A > A, B > B) but also emotions expressed in the other language (A > B, B > A). As an example of the latter situation, English listeners may be asked to recognize emotional prosody in Japanese, and vice versa. Although some studies (e.g. Albas et al. 1976, Dennis 1982, Gitter et al. 1972) used this reciprocal approach, the two cultural groups involved were

[18] Pell et al. (2009) report a significant in-group advantage but omitted the responses to one of the emotions ('neutral'). However, when the Pell et al. data are aggregated over all six emotional categories, there is no significant in-group advantage for the Argentinean Spanish listeners.

also ethnically different rather than just culturally-or-linguistically dissimilar. Other studies also adopted the reciprocal method (e.g. Ekman 1972) but only investigated the perception of facially expressed emotions between two culture groups. Even though previous studies have clearly indicated an in-group advantage in the cross-cultural perception of emotion, the reciprocal method was not used so that those studies present an incomplete picture, especially when it comes to the perception of vocal emotion. Therefore, I conducted the present study applying the reciprocal method to two cultural groups whose cultures and languages are distant from each other. In the present study, I chose Mandarin and Dutch listener groups for this research purpose. This is because Mandarin is a Sino-Tibetan tonal language which uses a wide pitch range (with pitch movements up to 12 semitones, Xu 1999), has monosyllabic words and a simple syllable structure (Duanmu 2007a, b), while Dutch is an Indo-European non-tone language, with a rather restricted pitch range (De Pijper 1983, 't Hart et al. 1990), with often long, polysyllabic (compound) words that may contain complex consonant clusters (Booij 1995). Moreover, Chinese and Dutch cultures are very much dissimilar, as one is Asian culture and the other is West-European culture. Therefore, the present study asks the following questions:

1) How well can Chinese and Dutch novice native listeners identify emotional prosody in Chinese and Dutch, and vice versa? In other words, what is the difference between the two listener groups in perceiving emotional prosody in their L1 and in an unknown language?
2) Is the cross-cultural perception of vocal emotion symmetrical between Chinese and Dutch, i.e., will Dutch and Mandarin listeners have similar abilities of identifying emotional prosody expressed in the other language?

My prediction for research question 2 is negative. The prosodic space which languages may use is finite. The parameters (or dimensions) of the phonetic space (and the prosodic space within it) can be used to express linguistic as well as paralinguistic contrasts. A functional principle holds that one can use a particular parameter in the phonetic space only once (e.g. Berinstein 1979, Potisuk et al. 1997, Remijsen 2002a, b). It follows from this functional principle that if a language uses a prosodic parameter for linguistic purposes, it can no longer use the same parameter for non-/paralinguistic uses – or, in a less extreme version of the theory – cannot use the same parameter as effectively for the expression of paralinguistic or extralinguistic meanings. As a consequence of the functional view, listeners of a tonal language will be less intent on (and in fact less experienced in) decoding the paralinguistic use of prosody than listeners of a non-tonal language. In other words, listeners of a non-tonal language are generally better at recognizing emotional prosody than listeners of a tonal language. Therefore, I would like to further predict that Dutch listeners would be overall better than Chinese listeners in correctly identifying emotional prosody in the present study.

In order to avoid terminological inconsistency I only use the term 'emotional prosody' in this chapter, and use it to refer to both vocally produced emotions (e.g. happiness, sadness, anger, fear, disgust) and attitudes (e.g. sincerity, irony, sarcasm). In this study six emotional prosodies have been studied: neutrality, happiness, (hot) anger, surprise, sadness and sarcasm.

### 7.2 Methods

Two perception experiments were conducted in the present study: the first perception experiment was designed to test how well Chinese native and novice Dutch listeners of Chinese perceive emotional prosody produced in Chinese. The second perception experiment aimed to find out how well the same two listener groups perceive the same emotional prosodies but portrayed in Dutch. This is the reciprocal situation: Chinese natives became the novice listeners of Dutch.

#### 7.2.1 Speakers

Four native Chinese speakers (2 males, 2 females, mean age = 45 years) whose mother tongue was standard Mandarin voluntarily took part in the recording of the stimuli for the first perception experiment. Four native Dutch speakers (2 males, 2 females, mean age = 33 years) voluntarily participated in the recording of the stimuli for the second perception experiment. These four native Dutch participants were also advanced L2 speakers of Chinese who learnt Chinese for 6 to 10 years and had been teaching Chinese as a foreign language for 2 to 10 years when the recordings were made. These speakers can easily switch between Dutch and Chinese. They translated the stimuli from Chinese to Dutch for the second perception experiment (see below). Most of the native Chinese and Dutch speakers had experience in stage performance in their mother tongue. Moreover, a mood induction technique in which different background stories were told to the speakers to express a stimulus in different emotions was applied.

#### 7.2.2 Listeners

Twenty native Mandarin listeners (10 males, 10 females, mean age = 24 years) and 20 native Dutch listeners (10 males, 10 females, mean age = 33 years) voluntarily participated in each of the perception experiments. The Chinese listeners were bachelor and master students at the University of Science and Technology Beijing who hailed from different parts of China; all spoke Mandarin. The native Dutch listeners were mainly bachelor students at the Humanities Faculty of Leiden University in the Netherlands and volunteers with variable education backgrounds. None of the native Dutch listeners spoke any Mandarin; neither did the native Chinese listeners speak any Dutch. The two listener groups were novice listeners of each other's native language (Chinese and Dutch). There was neither a special course in the curriculum nor any pre-test training designed for training these listeners to recognize emotions in their L1 or in an unknown language.

### 7.2.3 Materials and procedure

The testing materials for the two perception experiments were designed in a particular order. The stimuli were first collected and produced in Chinese and then translated into Dutch and portrayed in Dutch afterwards. The rationale behind this is that Chinese, a tone language, made the collection of stimuli more difficult, as the stimuli in Chinese had to meet the certain requirements (details see below). Dutch is a non-tonal language in which the requirements can be met with less difficulty. Therefore, it was easier to collect the stimuli in Chinese first and translate them into Dutch afterwards.

I selected six Mandarin statements as vocal stimuli (e.g. *She is three months pregnant; He has been to Xiao Ge's place once*) for the first perception test. The requirements for the stimulus selection are: (1) stimuli contain all the tones in Mandarin, i.e. 'high-level tone', 'rising tone', 'falling-rising tone', 'falling tone' and 'neutral tone' (e.g. Howie 1976); (2) stimuli have to be semantically neutral but can easily be expressed with different emotions; (3) both short and longer sentences have to be included, in case utterance length might play a role in the perception of emotional prosody.

According to the consensus of the native Chinese speakers and the Dutch L2 speakers of Chinese, the ensemble of six selected sentences met the requirements adequately. The stimuli were then translated into Dutch by the four Dutch L2 speakers of Chinese where sentence length, syntactic structure, syllables and sentence meaning were well controlled. In fact, some of the sentences may be associated more readily with some emotions than with others but on aggregate the lexico-syntactic materials will not be biased towards specific emotions. Each of the six statements was expressed in six different emotions (neutrality, happiness, anger, surprise, sadness and sarcasm). The list of stimulus sentences in Chinese is shown in Table 7.1.

Table 7.1. *Stimulus list in Chinese (Pinyin orthography) with English glosses.*

| | |
|---|---|
| 1. | **Shì nǐ.*<br>'It is you.' |
| 2. | *Xièxiè nǐ.*<br>'Thank you.' |
| 3. | *Xiǎo wáng wánquán bù zhīdào zhè jiàn shì.*<br>'Xiao Wang does not know about this matter.' |
| 4. | *Jīntiān xiàwǔ tā bùnéng lái cānjiā zhège huì.*<br>'He cannot attend the meeting this afternoon.' |
| 5. | *Tā huáiyùn sān ge yuèle.*<br>'She is three months pregnant.' |
| 6. | **Tā qùguò xiǎo gě jiā yì cì.*<br>'He has been to Xiao Ge's place once.' |

Note: '*': sentence was excluded in the second perception experiment. Macron '¯' = high-level tone, acute accent '´'= rising tone, haček 'ˇ' = falling-rising tone, grave accent '`' = falling tone; a syllable without tone mark has neutral tone.

Each of the six statements was expressed in six different emotions – neutrality, happiness, (hot) anger, surprise, sadness and sarcasm – by the four native Chinese speakers. The stimuli were digitally recorded (44.1 KHz, 16 bits) in a sound-proofed booth through a Logitech desk-top microphone. This procedure resulted in a stimulus set that consisted of 6 Chinese statements × 4 Mandarin speakers × 6 emotions = 144 emotional utterances.

For the second perception experiment, the four Dutch L2 speakers of Chinese were asked to express the same six emotional prosodies in Chinese. The stimuli were digitally recorded under the same conditions as in the first perception experiment. Two sentences were discarded from the stimulus set (see Table 7.1), as these two were less well perceived by the two listener groups in the first perception test. Therefore, the final stimulus set for the second perception experiment consisted of 4 Dutch statements × 4 Dutch L2 speakers of Chinese × 6 emotions = 96 discrete emotional utterances. The list of stimulus sentences in Dutch is shown in Table 7.2. It made the second experiment shorter than the first one. In the comparison between the two experiments, I only used the shared materials.

Table 7.2. *Dutch stimulus sentences with Broad IPA transcriptions and English glosses.*

| | |
|---|---|
| 1. | *Dank je wel.*<br>dɑŋk jə ʋɛl<br>'Thank you.' |
| 2. | *Xiaowang weet dat helemaal niet.*<br>ʃau ʋɑŋ ʋetɑt heləmal nit<br>'Xiao Wang does not know about this matter.' |
| 3. | *Vanmiddag kan hij niet naar de vergadering.*<br>ʋɑmɪdɑχ kɑni nit nar də ʋərɣadərɪŋ<br>'He cannot attend the meeting this afternoon.' |
| 4. | *Zij is drie maanden zwanger.*<br>zɛɪ ɪs dri mandə zʋɑŋər<br>'She is three months pregnant.' |

In both perception experiments, both of the listener groups including native Chinese and Dutch listeners were asked to make a forced choice of the speaker's intended emotion, from the six given emotions, immediately after they heard a stimulus. They also gave a confidence rating to each choice they made. A three-level confidence rating scale was used, with the following interpretation: 3 = 'The speaker expressed the intended emotion well. I am very confident of my answer', 2 = 'The speaker expressed the intended emotion moderately well. I am not so sure of my answer' and 1 = 'The speaker did not express the intended emotion well. I made the choice only by guessing.' The confidence scale was introduced in order to obtain a potential weighting factor such that responses given with great confidence would be weighted more heavily than responses that were largely based on guessing. The first experiment lasted 25 minutes and the second one lasted 15 minutes, including the time for the listeners to read the instructions in their native language before they started the test and a 6-second pause in between the emotional sentences for the participants to make a choice.

Each listener did the experiment individually in the presence of the experimenter. The stimuli were presented to the subject over closed headphones (but remained inaudible to the experimenter).

### 7.3 Results

The results proved insensitive to any weighting based on response confidence. Therefore, I report unweighted identification results only. Tables 7.3 and 7.4 are confusion matrices of intended versus perceived emotions in the two perception experiments by the two listener groups, i.e. native Chinese listeners and native Dutch listeners.

Table 7.3. *Perception of emotional prosody produced in Chinese by native Chinese speakers: Confusion matrix of intended and perceived emotions by Chinese (upper panel) and novice Dutch listeners (lower panel). Correct responses are located on the main diagonal (shaded).*

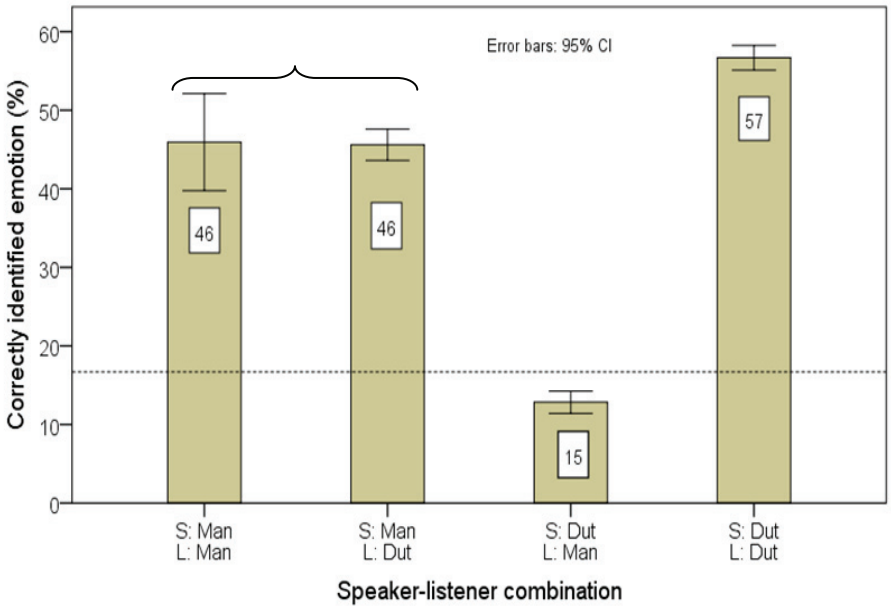
| Intended | Ang | Hap | Neu | Sar | Sad | Spr | Overall correct |
|---|---|---|---|---|---|---|---|
| | Responded emotion by Chinese native listeners | | | | | | |
| Angry | **56.3** | 4.8 | 10.2 | 5.2 | 10.0 | 13.5 | |
| Happy | 12.1 | **37.3** | 34.8 | 1.7 | .8 | 13.3 | |
| Neutral | 7.3 | 7.3 | **73.5** | 2.5 | 4.2 | 5.2 | |
| Sarcastic | 11.7 | 17.5 | 34.0 | **17.3** | 3.1 | 16.5 | 46.0 |
| Sad | 13.3 | 8.1 | 32.7 | 4.4 | **37.1** | 4.4 | |
| Surprised | 12.9 | 4.0 | 10.6 | 13.3 | 5.0 | **54.2** | |
| Total | 20.4 | 13.5 | 26.6 | 6.3 | 13.4 | 19.6 | |
| Responded emotion by Novice Dutch listeners | | | | | | | |
| Angry | **52.6** | 4.0 | 15.1 | 9.5 | 7.9 | 10.9 | |
| Happy | 35.7 | **20.4** | 14.1 | 5.8 | 3.6 | 20.4 | |
| Neutral | 4.0 | 4.2 | **71.2** | 9.9 | 7.3 | 3.4 | |
| Sarcastic | 13.3 | 6.5 | 15.7 | **28.8** | 16.1 | 19.6 | 46.0 |
| Sad | 5.2 | 2.4 | 28.6 | 10.5 | **49.2** | 4.2 | |
| Surprised | 9.9 | 12.3 | 5.0 | 6.3 | 15.1 | **51.4** | |
| Total | 19.1 | 7.4 | 25.8 | 10.4 | 17.8 | 18.6 | |

Table 7.4. *Perception of emotional prosody produced in Dutch by native Dutch speakers: Confusion matrix of intended and perceived emotions by native Dutch (upper panel) and novice Chinese listeners (lower panel). Correct responses are located on the main diagonal (shaded).*

| Intended | Ang | Hap | Neu | Sar | Sad | Spr | Overall correct |
|---|---|---|---|---|---|---|---|
| | Responded emotion by Dutch native listeners | | | | | | |
| Angry | **55.6** | 9.5 | 9.0 | 12.5 | 7.9 | 3.7 | |
| Happy | 2.3 | **44.9** | 4.6 | 17.4 | 1.2 | 12.5 | |
| Neutral | 15.9 | 13.0 | **68.3** | 17.6 | 19.7 | 5.3 | |
| Sarcastic | 13.4 | 7.9 | 3.9 | **34.3** | 6.3 | 6.0 | 57.0 |
| Sad | 8.1 | 9.7 | 13.4 | 4.4 | **64.6** | .0 | |
| Surprised | 5.6 | 15.0 | .7 | 13.9 | .5 | **72.5** | |
| Total | 16.3 | 13.1 | 20.0 | 10.0 | 18.9 | 21.2 | |
| | Responded emotion by Novice Chinese listeners | | | | | | |
| Angry | **8.0** | 17.8 | 32.8 | 5.5 | 20.0 | 16.0 | |
| Happy | 13.0 | **8.2** | 36.5 | 16.0 | 13.5 | 12.8 | |
| Neutral | 31.0 | 7.2 | **22.2** | 11.0 | 14.0 | 14.5 | |
| Sarcastic | 18.0 | 8.8 | 33.2 | **11.8** | 9.2 | 19.0 | 15.0 |
| Sad | 17.5 | 12.0 | 26.8 | 13.2 | **21.2** | 18.8 | |
| Surprised | 16.7 | 10.2 | 30.4 | 11.2 | 14.3 | **17.2** | |
| Total | 8.9 | 9.1 | 24.7 | 13.1 | 23.6 | 19.1 | |

The confusion matrices and Figure 7.1 together show that Dutch native listeners were able to recognize the vocal emotions well above chance level (= 16.7%) regardless the speaker type; in contrast to this, Chinese native listeners were only able to identify the emotional prosodies produced in their native language reasonably well, but failed to recognize them above chance level when the emotions were vocally portrayed in the unknown language (Dutch). Table 7.3 and Figure 7.1 indicate that Dutch novice listeners perceived the six emotional prosodies produced by native Chinese speakers as well as Chinese native listeners, with a mean correct identification rate of 46%. Moreover, Figure 7.1 shows that Dutch native listeners recognized the same six emotional prosodies produced in their native language (Dutch) substantially better than in Chinese, with a mean correct recognition rate of 57%. It can be also seen from Figure 7.1 and Table 7.3 that Chinese native listeners identified the emotional prosodies portrayed in their L1 significantly better than in an unknown language too, with a mean correct recognition rate at 46% in the first perception experiment and at 15% in the second perception experiment. These findings support previous studies, e.g., Elfenbein & Ambaby (2002), Pell et al. (2009), Thompson & Balkwill (2006), claiming that listeners generally recognize emotional prosody produced in their L1 better than in an unknown language, which phenomenon is known as the 'in-group advantage'. In addition, Van Bezooijen's (1984) study shows that Dutch native listeners were significantly better than Asian listener groups (Chinese and Japanese) in recognizing emotional prosody in Dutch, which is along the line of the present finding. The present result also shows that Dutch native listeners could identify vocal emotion in an unknown language equally well as native listeners of that language could. These findings imply that some culture group (e.g., Dutch) might be generally better in recognizing vocal emotion than some

other culture group, for example, Chinese, no matter whether vocal emotion is produced in the listeners' L1 or in an unknown language. In other words, it might be the case that perception of emotional prosody cross-culturally is not symmetrical.




*Note: 'S' = speaker type; 'L' = listener type; 'Man' = native Mandarin speaker; 'Dut' = native Dutch listener. Chance level = 16.7%, which is marked by the dotted line in the graph.

Figure 7.1. *Percent correct identification (%) of emotions for four combinations of speaker (S) and listener (L) types.* Means and 95%-confidence limits are indicated. Conditions under the same brace do not differ significantly from each other by a Bonferroni post-hoc test* ($\alpha = .05$).

The finding that Chinese native listeners were not able to identify (six) emotional prosody produced in Dutch above chance (= 17%) is not in line with the finding of Van Bezooijen (1984). In her study, novice (Taiwanese) Chinese listeners were able to recognize (ten) Dutch emotional prosody well above chance level (= 10%). The mean correct identification rate for the (Taiwanese) Chinese was 37%, which however, was significantly lower than that of the Dutch native listeners (66%). The contradictory finding might result from the varieties of the native Chinese listeners' dialectical backgrounds, as the listeners hailed from all over mainland China which covers at least ten official dialects (Li 1987, a-2).

Surprisingly, Table 7.3 indicates that native Chinese and novice Dutch listeners followed a rather similar recognition order, such that they both found 'neutrality' the easiest emotion to identify, followed by 'anger', 'surprise', 'sadness', 'happiness' or 'sarcasm'. Table 7.4 shows the perception of the same emotional prosodies produced in

Dutch by the two listener groups. The detailed recognition order of the six emotional prosodies by the two listener groups in the two perception tests is shown in Table 7.5. In the second perception experiment, Dutch native and Chinese novice listeners did not follow any similar recognition pattern. It implies that the two culture groups resort to different cognitive resources when perceiving the vocal emotions. The concrete cognitive resources can hardly be known at this stage.

Table 7.5. *Recognition order of the six emotional prosodies by native Chinese and Dutch listeners. The recognition order of the six Chinese emotional prosodies is presented in the upper panel; the recognition order of the six Dutch emotional prosodies is presented in the lower panel.**

| Listener group | Recognition order of the six Chinese emotional prosodies | | | | | |
|---|---|---|---|---|---|---|
| Native Chinese | neutrality > | anger > | surprise > | happiness > | sadness > | sarcasm |
| Dutch novice | neutrality > | anger > | surprise > | sadness > | sarcasm > | happiness |
| Listener group | Recognition order of the six Dutch emotional prosodies | | | | | |
| Native Dutch | surprise > | neutrality > | sadness > | anger > | happiness > | sarcasm |
| Chinese novice | neutrality > | sadness > | surprise > | sarcasm > | happiness > | anger |

*: '>' means 'better identified than'.

Table 7.5 shows that both Chinese and Dutch native listeners perceive emotional prosody differently in their L1 and in the unknown language in terms of recognition order. It seems that negative emotions, such as sadness or anger, are relatively easy to recognize for both listener groups. This finding supports some previous studies, e.g., Ohman et al. (2001), Tooby & Cosmides (1990), claiming that negative emotions are generally better recognized by human beings. Even though anger is generally considered a basic emotion, the Chinese novice listeners massively confused 'anger' with 'neutrality' for Dutch emotional prosody. It implies that perception of primary emotions (e.g., anger or neutrality) only through the audio-channel is not universal all the time. It confirmed that perception of emotional prosody is partly universal and partly culture-or-language specific, even for some basic emotions, for example, anger, neutrality or sadness. In some extreme case, it is possible that listeners are not able to identify emotional prosody in an unknown language above chance, especially the unknown language is typologically remote from the listener's L1.

Finally, I will analyze the confidence ratings. Although it was mentioned earlier in this section that there was no effect of weighting on the results and therefore only unweighted identification results were presented, I would like to make use of the confidence ratings all the same to investigate the social behaviour of the listener groups.

In this case, I just present means and observe unexpected differences between the groups.

Figure 7.2A-B shows that Chinese native listeners were less confident than the Dutch listeners in their identification of vocal emotions in their L1. Surprisingly, however, Chinese novice listeners were more confident than Dutch natives themselves when recognizing Dutch emotional prosody. The Chinese listeners showed less confidence when responding to emotions produced in their own language (mean = 1.5) than when responding to the Dutch emotions (mean = 2.3). The difference is highly significant by a paired-samples t-test, $t(19) = -39.1$ ($p < .001$). The same statistical method was applied to the Dutch listeners' recognition of emotional prosody in Chinese and Dutch. The result shows that there was also a significant effect for the Dutch listeners in the two perception experiments, $t(19) = -7.5$ ($p < .001$), indicating that Dutch listeners are more confident in identifying vocal emotion both in their L1 than in the unknown language.

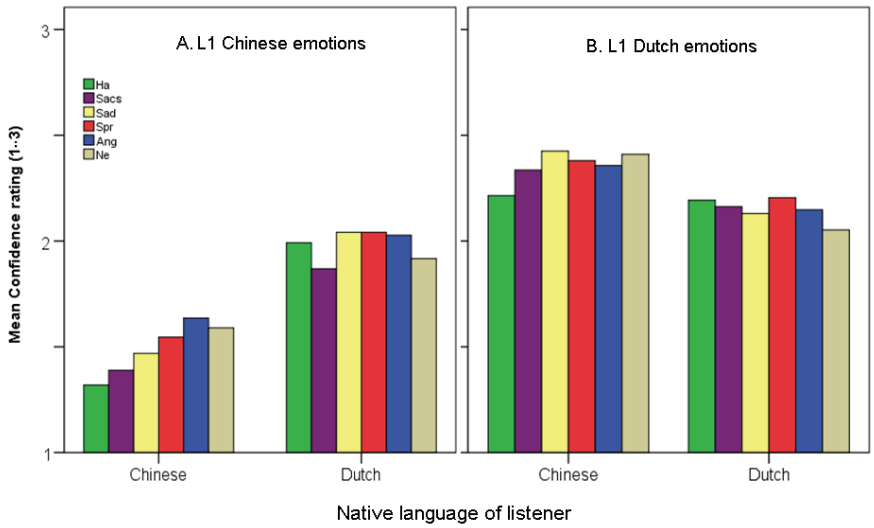



Figure 7.2. *Confidence rate (3 = most) of six intended emotions by native Chinese and Dutch listeners. Panel A presents the ratings of emotional prosody produced by native Chinese speakers. Panel B presents rating of emotional Dutch prosody encoded by advanced Dutch L2 speakers of Chinese.*

## 7.4 Conclusions and discussion

The results of this investigation indicate that Chinese native listeners are only able to recognize emotional prosody in their native language reasonably well, but cannot identify emotions in an unknown language above chance level, meaning that their identification in the unknown language is based on guessing. In contrast to this, Dutch native listeners are able to identify vocal emotion in an unknown language as accurately

as native listeners. And they recognize the same emotional prosodies expressed in their L1 even more correctly. Moreover, 'anger' was perceived relatively successfully by the two listener groups, apart from being identified by the Chinese listeners in Dutch. These findings show that perception of emotional prosody can be partly culture or language specific and partly universal. It confirms previous studies which claim that perception of vocal emotion is a combination of universal elements and cultural or linguistic variables (for example, Pell et al. 2009, Scherer 2000, Van Bezooijen (1984). The results show a significant difference between listeners' identifications of vocal emotions in their L1 and in the unknown language. Therefore, the existence of the in-group advantage found by other researchers (e.g. Elfenbein & Ambaby 2002, Pell et al. 2009, Thompson & Balkwill 2006) is confirmed, which claims that listeners generally better recognize emotional prosody produced in their L1 than in an unknown language. Since the Dutch novice listeners recognized the Chinese emotional prosodies as well as the Chinese natives did, therefore, the 'in-group advantage effect' might not be developed extensively, predicting that native listeners are generally better than novice listeners in perceiving vocal emotions in the native language. In fact, for some cultural groups (e.g. Dutch), novice listeners can identify vocal emotion in unknown native language as well (and as confidently) as native listeners can (or even better).

Chinese native and Dutch novice listeners followed a rather similar recognition order for the perception of Chinese vocal emotions. This finding is in line with findings by Scherer et al. (2001), who reported that patterns of confusion were very similar across all cultural groups: 'These data suggest the existence of similar inference rules from vocal expression across cultures' (Scherer et al. 2001). However, the confusion patterns of the two listener groups' perception of Dutch vocal emotions differed substantially in the present study, implying that there is an overlap between Dutch and Chinese inference rules from vocal expression; however, Dutch inference rules might cover a relatively wider range than those of Chinese. My results further suggest that perception of emotional prosody is not symmetrical cross-culturally, such that listeners of two barely related cultural or linguistic groups might have different perceptual abilities of inferring vocal emotions expressed in the other language.

A peripheral finding of the present study is that Chinese listeners are more confident in identifying vocal emotion in an unknown language than in their L1, even though their identifications of emotional prosody in the unknown language seemed mainly based on guessing. This unexpected social behavior of Chinese listener group cannot be explained at the current stage. In contrast to this, the mean confidence rating by Dutch listeners increased when the listeners were asked to perceive the vocal emotions in their L1. This finding can be seen as a spin-off of the in-group advantage, predicting that listeners are generally more confident at identifying vocal emotion in their L1 than in an unknown language. However, this prediction might not apply to all the cultural groups, for example, Chinese. It needs to be further tested by involving more cultural groups, especially cultural groups which culturally differ from each other very much.

The asymmetrical performance between Chinese and Dutch in the two perception experiments forces us to answer research question 2 negatively. In the asymmetrical approach one cultural group may be generally more sensitive to correctly identifying

emotional prosody than some other cultural group, e.g. Chinese. This view needs to be further tested by involving more and more diverse cultural groups. Moreover, Scherer et al. (2001) found that, generally, accuracy of identification of vocal emotions by different cultural groups decreased with increasing language dissimilarity between the different cultural groups. It is concluded that culture- and language-specific paralinguistic patterns may influence the decoding process. Up to this point, it can be true that Chinese emotional prosody generally contains more universal cues for novice listeners to detect. In contrast to this, Dutch emotion prosody may bear more culture-or-language-specific variables which are not easy for novice listeners to decode. This would explain why Chinese native listeners did poorly when instructed to identify emotional prosody in Dutch. But this does not explain the asymmetry in the success with which Dutch and Mandarin listeners identified the emotional prosodies in the other language. It has been suggested that such asymmetries may arise as a consequence of either cultural differences or differences between the phonologies of the two languages involved. McCluskey et al. (1975) consider, on the basis of the asymmetry they found, that Mexican-Spanish children, in contradistictions to their (Canadian) English peers, are brought up in a cultural setting that emphasizes the importance of attending to emotional prosody. This would then explain why Mexican children identify emotional prosody better than Canadian-English children, not only in their own language (Mexican Spanish) but also in Canadian English. But we may also think of strictly linguistic reasons that could explain asymmetries in affect perception, especially in the comparison of Dutch and Mandarin.

Apparently, our prediction made earlier that Dutch listeners would be overall more sensitive than Chinese listeners in correctly identifying emotional prosody was supported by the results. Therefore, it implies that the functional view might be true. If a language uses a prosodic parameter for linguistic purposes, it can no longer use the same parameter for non-/paralinguistic uses; or cannot use the same parameter as effectively for the expression of paralinguistic or extralinguistic meanings. In addition, Ross et al. (1986) have shown there is less use of short-term changes in F0 to express emotion in tone languages (in which short-term F0 contours are used to carry lexical information) than in Indo-European languages (in which F0 plays no lexical role). Thus it seems that, in some cases at least, use of a particular acoustic feature in spoken language limits its use for the communication of emotion. This insight can be incorporated into the functional view, indicating that lexical tone might suppress or inhibit the expression of emotional prosody. In the present case, Mandarin is a tonal language that uses a wider pitch range than Dutch, and does so for linguistic purposes. Accordingly, its prosodic space for emotion is smaller than that of Dutch. As a consequence of the functional principle, listeners of Chinese are, in fact, less experienced in decoding the paralinguistic use of prosody than listeners of Dutch. This would explain why the Chinese listeners performed no better than the Dutch novices for the Chinese emotional prosodies and why they did so poorly in identifying Dutch vocal emotion, compared to the natives.

The unexpected performance difference between the Chinese and Dutch listeners in the two perception tests might be enhanced by the absence of particles in the Chinese stimuli used in the experiment. In everyday Chinese speech particles often appear at the

end of a sentence, carrying considerable emotional information that is alternatively expressed by intonation in other languages.[19] Since this kind of lexical markers were deliberately left out in the present study due to the research purpose, their absence might have affected the perception of the emotional prosodies by L1 listeners but not by the Dutch listeners. Moreover, Dutch listeners might generally be better equipped for extracting emotional meaning from prosody, according to the functional view. That is possibly why the Chinese L1 listeners did not outperform the Dutch listeners in the perceptual study in which the stimuli with no final particles attached were presented through the audio channel only. Testing this hypothesis is beyond the scope of the present study. Part of the endeavor would be to determine how much use Mandarin and Dutch make of particles expressing emotions on the part of the speaker and what the division of work would be between the use of such particles and emotional prosodies.

[19] Lexical markers here refer to final particles in Chinese which can carry emotional information. Examples would be *ya* (friendly) or *a* (enthusiastic). Syntactic markers may be used imply negative emotions such as the 'annoyance' marking construction *nán dào … (ma)?*, which is a rhetorical question confronting the listener with his/her ignorance (less negative with *ma* than without).

# Chapter Eight

# Conclusion and Discussion

## 8.1 Introduction

In this dissertation, I have investigated 1) the perception of L1 and L2 produced Chinese emotional prosody by natives, naïve Dutch listeners and advanced learners of Chinese; 2) how well Dutch L2 speakers of Chinese vocally produced emotions in the L2 and how well they expressed the same emotions in their L1 (Dutch); 3) the in-group advantage and different cultural groups' abilities of identifying vocal emotions in an unknown language. In order to answer all the research questions listed in Chapter 1, three judgment studies were conducted in this dissertation. The first judgment study was designed to test how well native Chinese listeners, naïve Dutch listeners and advanced learners of Chinese perceive Chinese emotional prosodies produced by native Chinese speakers. The results were used as the baseline for comparisons. The second judgment study aimed to find out how well the same three listener groups perceive the same Chinese emotional prosodies produced by the Dutch L2 speakers of Chinese. The production of the Chinese emotional prosody by the Dutch L2 speakers was also investigated in this judgment study. The third judgment study was carried out in a reciprocal way to test the in-group advantage and the Chinese and Dutch native listeners' abilities of identifying emotional prosody in their L1 and in the other language. In this chapter, I will summarize the main findings of the three judgment studies and give the possible explanations of some unexpected results, as well as to show which direction future studies can go.

## 8.2 Answers to research questions

In this section, I will recapitulate the research questions formulated in Chapter 1 and provide integrated answers to them.

### 8.2.1 Perception of native-Chinese emotional prosody by three listener groups

Three listener groups participated in the first judgment study, i.e. native Chinese, Dutch naïve listeners and advanced Dutch learners of Chinese. The first judgment study was designed to test how well the three listener groups perceive Chinese emotional prosodies vocally expressed by Chinese native speakers, which is research question one. This study also aimed to map out the confusion patterns of the listener groups, which

addresses a sub-question of the research question one. The results were used as the control condition for the second judgment study.

The results show that the three listener groups were able to identify the six Chinese emotional prosodies above chance level (chance level = 16.7%). The Dutch L2 learners of Chinese performed best in the perception experiment (with the mean recognition rate 54%) and native Chinese and Dutch naïve listeners performed equally well (with the mean recognition rate 46% in both cases). These results show that Dutch naïve listeners could recognize emotional prosody in an unknown language as well as native listeners of that language (Chinese); Dutch L2 learners of Chinese identified the Chinese emotion prosodies better than naïve Dutch listeners and therefore also better than the native Chinese listeners themselves. This finding is compatible with earlier results by Shoshi and Gagnié (2010) showing that experienced L2 learners of a language are better than listeners with no experience of the L2 in perceiving vocal emotions in the L2. Detailed confusion patterns can be seen from Table 3.2. Basically, 'neutrality' was identified correctly most often among all the emotions by the three listener groups, followed by 'anger' and 'surprised'. 'Sarcasm' was not well identified by all the listener groups, implying that some non-primary emotions, e.g. 'sarcasm', are more individual-specific. Chinese native and Dutch naïve listeners shared some similar confusion patterns; however, the confusion patterns of Dutch L2 learners of Chinese were in between of those of the Chinese native and Dutch naïve listeners. This shows that L2 learners of a target language have a hybrid system when perceiving emotional prosody produced in the target language. This hybrid system is partly influenced by speakers' L1 and partly affected by the target language itself. This would in return explain why the confusion matrix of the Dutch advanced learners of Chinese was in-between of the matrices of the native Chinese and the Dutch naïve listeners in the first judgment study.

### 8.2.2 Perception of L2 Chinese emotional prosody by three listener groups

The second group of research questions is: how well can native Chinese, Dutch naïve listeners and advanced Dutch learners of Chinese perceive the six Chinese emotional prosodies vocally portrayed by the Dutch L2 speakers of Chinese? What will be the confusion patterns of the three listener groups? Will the confusion patterns be similar to those in the perception experiment where the emotional prosodies were vocally portrayed by native Chinese speakers? In order to answer these questions, the same listener groups, i.e. native Chinese, naïve Dutch listeners and advanced Dutch learners of Chinese, participated in the first perception experiment of the second judgment study, in which they perceived Chinese emotional prosodies produced by the L2 speakers of Chinese.

The results show that the three listener groups could not perceive the same six Chinese emotional prosodies vocally portrayed by L2 speakers of Chinese as well as they did with those produced by native speakers. The mean recognition rates of the three listener groups are 39% with Chinese native listeners, 38% with Dutch naïve listeners and 41% with advanced Dutch learners of Chinese. Even though the advanced Dutch learners of Chinese performed slightly better than the other two listener groups, there was no significantly better group in the perception test, which means that the three

listener groups did equally well/poor. It also implies that L2 speakers produced emotional prosodies that are less recognizable than those expressed by natives. The detailed confusion patterns of the three listener groups can be seen in Table 4.3. Admittedly, there are similarities between the confusion patterns obtained from the Chinese native and the Dutch naïve listeners in the perception of the non-native produced emotional prosodies. However, the confusion patterns are not very similar to those that were obtained in the perception experiment where the emotional prosodies were vocally portrayed by native Chinese speakers. It implies that the native Chinese speakers and the L2 speakers of Chinese may have used different vocal correlates to express the emotions. In the perception of the L2-produced Chinese emotional prosodies, the confusion categories which advanced Dutch learners of Chinese fell into are quite similar to those of naïve Dutch listeners. For example, they showed the exact same tendency as naïve Dutch listeners for 'sarcasm': they often misidentified 'sarcasm' as 'neutrality' (19.7%) and 'surprise' (19.7%); and naïve Dutch listeners misrecognized it with 'neutrality' (18.4%) and 'surprise' (18.4%). At this point, one can assume that L1-transfer is an important strategy in interpreting paralinguistic meaning (e.g. emotional prosody) in L2. However, the hybrid system seen in the first judgment study did not apply to the advanced Dutch learners of Chinese in the perception of the L2-produced Chinese emotional prosodies. It implies that advanced learners of a target language may mainly resort to their native language to perceive emotional prosody expressed by L2 speakers from the same linguistic group.

Overall speaking, the difference between the confusion matrices of the three listener groups in the perception of native and non-native produced Chinese emotional prosodies is big. It implies that native and non-native speakers may have used very different acoustic correlates to produce emotional prosody in the target language. It also indicates that perception of non-native produced emotional prosody is more language-or-culture specific.

The confidence scale was originally introduced to obtain a potential weighting factor such that responses given with great confidence would be weighted more heavily than responses that were largely based on guessing. It turned out that the results of the first and second judgment studies proved insensitive to any weighting based on response confidence. However, I would like to make a use of the confidence rating all the same to look tangentially at the social behavior of the three listener groups in the two judgment studies. The results show that Chinese native listeners were more confident in perceiving emotional prosodies produced by L2 speakers of Chinese. The reason for this behavior is not clear since one would expect listeners to be more confident when having to make decisions based on materials produced by speakers who share the same linguistic code. In contrast to the above, both of the naïve Dutch listeners and the advanced Dutch learners of Chinese were confident in the perception of Chinese emotional prosody produced by both natives and non-native speakers.

### 8.2.3 Production of emotional prosody in Dutch L2 speakers' L2 and L1

The second judgment study had two aims: one is to test whether L2 speakers of Chinese are able to produce emotional prosodies in the L2 as well as they do in their L1; the other is to find out the similarities and differences between the two productions. Twenty native Dutch listeners identified the same six emotional prosodies produced by the same L2 speakers of Chinese in their native language – Dutch. The results show that the mean correct recognition rate of the native Dutch listeners increased significantly when the emotions were vocally produced in speakers' L1. It means that the Dutch L2 speakers of Chinese are better at vocally expressing emotions in their native language. According to the combination study of the first and second judgment studies in Chapter 5, we can see that the Dutch L2 speakers of Chinese were not able to vocally produce emotion very successfully as native Chinese did, even though their proficiency of Chinese is very high. Neither were they able to express emotional prosody in their L2 as well as in their L1. Therefore, we can conclude that L2 limits L2 speakers' vocal expression of emotions, especially when L2 speakers' L1 is not a tonal language but their L2 is.

The confusion matrices suggest that the advanced L2 speakers of Chinese have developed an in-between manner of producing emotional prosody in the L2. This in-between manner is neither very much similar to the production of Chinese native speakers nor completely like the manner they use to produce emotional prosody in their L1 (Dutch). There are more explanations about this in-between manner in § 8.2.6 in this chapter, which is named as 'the hybrid system'. Moreover, the results also suggest that L2 speakers' ability of successfully producing emotional prosody in the L2 cannot be obtained automatically during the learning process of the L2 in general. It seems that this ability does not go along with the increasing of one's language proficiency in the L2. It is not clear at this stage whether this ability can be trained by designed curriculums. This needs to be tested in practice.

### 8.2.4 Lexical tone and expression of emotional prosody

At the beginning of the present study, I predicted that if a language uses a prosodic parameter for linguistic purposes, it will have less space for non-/paralinguistic use of the same parameter. If this prediction were true, speakers of a lexical tone language (such as Mandarin) should have less room to express emotion through prosody (specifically through paralinguistic use of speech melody) than speakers of a non-tone language (such as Dutch or English). In a more extreme version, it can be interpreted such that listeners of a non-tonal language are generally better in perceiving emotional prosody than listeners of a tonal language. Interestingly, the results of the first and second judgment study showed that naïve Dutch listeners, whose L1 is a non-tonal language, performed equally well as Chinese native listeners, whose L1 is a tonal language. Moreover, the L2 learners of Chinese performed even better than native Chinese listeners did.

In addition, a perception experiment designed in a reciprocal way was carried out later, in which Chinese and Dutch novice listeners perceived the six vocal emotions

expressed in their native language and in the other language. The results showed that Chinese listeners were only able to identify emotional prosody expressed in their L1 (Chinese) reasonably well but failed to recognize the same emotional prosody portrayed in an unknown language (Dutch) above chance level (= 16.7%). This means that Chinese novice listeners of Dutch are not able to identify emotional prosody in an unknown language (Dutch); they can only recognize emotional prosody in their L1 well. On the contrary, Dutch novice listeners of Chinese could recognize emotional prosody expressed in an unknown language (Chinese) reasonably well (mean recognition rate = 46%); and they identified emotional prosody portrayed in their native language (Dutch) significantly better than they did with the unknown language – Chinese. These findings together support the prediction made earlier – that listeners of a non-tonal language are generally better at perceiving emotional prosody than listeners of a tonal language. Furthermore, this prediction can also explain why L2 speakers of a tonal language are not able to express emotional prosody as well as they do in their L1. If a tonal language limits both L1 and L2 speakers of the language in the production of emotional prosody, it would limit L2 speakers more. Because L1 speakers of the tonal language can automatically get the lexical tones correctly while producing emotional prosody, but L2 speakers of the tonal language cannot do it automatically. In other words, when L2 speakers of the tonal language try to vocally portray emotions in the tonal language, they may have to pay much attention to getting the lexical tones right while portraying the emotion at the same time. It will result in that they have even lesser room than native speakers for paralinguistic uses, as they have to make lexical tones right first consciously or unconsciously and then put emotional prosody on top of the lexical tone afterwards. It is very difficult for L2 speakers of a tonal language to do these two things at the same time, even though they have very high proficiency of the L2. Therefore, the Dutch L2 speakers did not portray the emotional prosodies as well as they did in their L1. In summary, the prediction is a sensible explanation for the results. But it needs to be tested more extensively, for example, it can be tested with British naïve listeners and British L2 learners of Chinese; or German naïve listeners and German learners of Chinese perceiving Chinese emotional prosody and vice versa.

### 8.2.5 Acoustic correlates of emotions: recognition by humans and machine

In Chapter 6, I examined the value of eight parameters as correlates of the six emotions studied. The eight parameters were the same for each of the three groups of speakers, i.e. Mandarin L1, Mandarin L2 and Dutch L1 (the latter two were the same individuals). The acoustic analysis shows that fundamental frequency, including mean F0, SD_F0 and 'slope of the F0', is an influential variable in the production of vocal emotions by the three groups of speakers. This finding confirms the study of Scherer (1996), who claimed that F0 plays a crucial role in the production of emotional prosody. However, jitter and standard deviation of the intensity did not contribute much to differentiating between emotions in the present study. Never were more than two subgroups of the emotional prosodies differentiated for any of the three speaker groups. This finding contradicts to the traditional claims, indicating that jitter and intensity play an important role in the production and perception of vocal emotions (e.g. Bachorowski 1999, Biersack & Kempe 2005, Scherer 1996). The acoustic analysis also shows that 'tempo' and 'compactness' were only sensitive to Mandarin L1 speakers, for whom three

subgroups of the emotional prosodies were found. Slope of the F0 indicates that Chinese uses rising intonation to express surprise, which confirms the previous studies (e.g. Yip 2006), claiming that many tonal languages use rising intonation to express surprise. Moreover, HNR was only relatively sensitive to Mandarin L2 speakers but not to Dutch and Mandarin L1 speakers. These altogether further show that fundamental frequency is a very influential variable in the production and perception of vocal emotion in general; however, other acoustic factors are not universally important; they are more language-specific or emotion-specific.

The acoustic analysis indicates that some basic emotions such as 'happy' and 'angry' can be clearly discriminated from each other by mean F0 and SD_F0, regardless the speaker type. 'Happy' is characterized by high values for mean and SD of F0 (z-values close to 1) while 'angry' has z-values close to 0. 'Neutral' is also universally differentiated from 'happy' and 'angry', viz. by low values for mean and SD of F0 (z-values close to –1). However, more controlled emotions, e.g. 'surprised' and 'sarcastic', are not well classified by any of the eight parameters examined in Chapter 6. HNR can clearly distinguish 'sad' from 'neutral' with Mandarin L2 and Dutch L1, who were actually the same individuals, but not in the case of L1 Mandarin speakers.

Furthermore, some other factors can also influence the perception of vocal emotion, for instance, personal interpretation of the emotional label. Since 'surprised' includes both positive and negative surprise, the human listeners sometimes misinterpreted this emotion as 'happy' or 'angry', respectively. In addition, some emotions were expressed differently by male and female speakers within one particular language. Moreover, the acoustic analysis shows that vocal emotions cross-culturally were produced diversely too. Therefore, we can conclude that production and perception of vocal emotion by humans is a much more complex and integrated procedure. It involves not only acoustic correlations but also other factors, such as, sex, language or personal interpretation of the emotional label. This conclusion is in the line with the previous studies, claiming that other factors including talker sex, talker linguistic background, talker identity and personal interpretation of emotional label may well also prove to be important in the production of vocal emotions (Bachorowski 1999, Scherer 2003).

In Chapter 6, I also conducted the automatic recognition (Linear Discriminant Analysis – LDA) of the six emotional prosodies. The results of LDA show that the automatic recognition in the present study can recognize human-produced emotional prosody well above chance level (50% overall correct). There was significant correlation between confusions obtained by the automatic recognition and by the human listeners in the present study. Moreover, the overall recognition rate of LDA is slightly better than that of the human perception. This indicates that automatic recognition can reflect human perception of emotional prosody to some extent; however, the human perception is still different from the computer perception. There are still acoustic correlates which used by the algorithm to discriminate between emotions but not used by L1 and L2 listeners in reality. In addition, the Stepwise LDA shows that there are four parameters which significantly contribute to the production and perception of emotional prosody: 'utterance duration', 'fundamental frequency', 'compactness' and 'HNR'. There may also be some other acoustic parameters contributing to the

production and the perception of emotional prosody in general, which have been missed in this dissertation.

### 8.2.6 L1-transfer and the hybrid system

The acoustic analysis indicates that Dutch L2 speakers use some acoustic parameters in the production of emotional prosodies in the L2 (Chinese) the same way they do in their L1 (Dutch), e.g. standard deviation of F0 and standard deviation of intensity. Therefore, we may conclude that L1-transfer is a strategy for L2 speakers to vocally produce emotions in the L2. However, this strategy may not work for all the emotions, e.g. not for 'surprise' and 'sarcasm'. Moreover, the results of the acoustic analysis show that the advanced L2 speakers of Chinese have developed a hybrid system of producing emotional prosody in the L2. This (L2) hybrid system approximates to some extent the Chinese native manner of portraying vocal emotion (the way it involves utterance duration, mean F0, slope of the F0, compactness and jitter), but exploits the variability in F0 and intensity that the L2 speakers use to produce emotional prosody in their L1. Emotional prosodies produced in this in-between manner were identified above chance level by both the native and non-native listeners in the present study. However, these emotional prosodies are less recognizable overall (41% correct within-group identification) than those produced in the Chinese native manner (46% correct). This would indicate that the expression of emotion through prosody is limited in an interlanguage. It further supports that L2 limits the expression of emotional prosody. The results in the present study also suggest that the L2 speakers did not automatically acquire the native approach to vocally produce emotional prosody in the target language during their L2 learning process. Therefore, it seems that, in a situation where there is no particular training for how to produce emotional prosody in L2, L2 speakers will create their own hybrid system/manner to express vocal emotions in the L2. This hybrid solution works for some emotions but not for all. Furthermore, this hybrid solution was also seen in the perception of native-produced Chinese emotional prosody by Dutch advanced learners of Chinese (see § 8.2.2). It implies that this hybrid system/solution may be applied in both of the production and the perception of emotional prosody in an L2.

In fact, the hybrid system can be seen as an extension of the interlanguage theory in terms of paralinguistic communication in L2. The interlanguage theory claims that L2 learners of a target language who have not become fully proficient yet but are approximating the target language, will preserve some features of their first language (or L1) or overgeneralize target language rules in speaking or writing the target language and creating innovations (Selinker 1972). In the present study, we have seen the same phenomena happen in the production and the perception of emotional prosody in Chinese by Dutch L2 advanced learners of Chinese. It remains to be seen to what extent the learning of emotional prosody can be modelled with rules and to what extent the concepts of the interlanguage theory apply to paralinguistics.

### 8.2.7 In-group advantage and cross-cultural perceptual ability of vocal emotion

The third judgment study conducted in a reciprocal way was designed to find the answer to the sixth research question, which is whether the in-group advantage found by other researchers is universal, claiming that listeners are better at recognizing emotional prosody produced in their native language than in their L2 or an unknown language. Twenty Chinese and 20 Dutch native listeners who did not know any Dutch and Chinese (respectively) perceived the emotional prosodies in their L1 and in the other language. In this case, Dutch was the unknown language for the Chinese native listeners; conversely, Mandarin Chinese was the unknown language for the Dutch native listeners. In the third judgment study, all the emotional prosodies were expressed by native speakers of Chinese and Dutch. The results show that both Chinese and Dutch native listeners recognized emotional prosodies better in their native language than in the unknown language. This finding suggests that the in-group advantage might be universal – but only in this weaker sense; it does not necessarily mean that vocal emotions are always recognized best by native listeners of the language the emotions are produced in. As a case in point, Dutch novice listeners identified the Chinese emotional prosodies as well as the natives did – and recognized the emotional prosody in their L1 much better. In contrast to this, Chinese native listeners identified the emotional prosody in their L1 reasonably well but failed to recognize the vocal emotion expressed in the unknown language (Dutch). This indicates that Dutch and Chinese listeners have very different abilities of identifying emotional prosody in an unknown language. It means that some culture groups might be generally better than some other culture groups in identifying emotional prosody. This further supports the view that cross-cultural perception of emotional prosody is not necessarily symmetrical.

### 8.2.8 Universal vs. culture-or-language specific

The ultimate goal of this dissertation is to find the answer to the last research question: whether perception and production of vocal emotion are universal or rather more language-specific and/or culture specific. The first two judgment studies tested how well native Chinese, naïve listeners and advanced learners of Chinese perceived Chinese emotional prosodies produced by native and L2 speakers; and how well L2 speakers of Chinese produced emotional prosody in the L2 and in their native language. The third judgment study investigated how well Chinese and Dutch novice listeners perceive emotional prosodies expressed in their L1 and in the other language. The three judgment studies all show that both perception and production of emotional prosody are partly universal and partly language-or-culture specific.

Perception of some basic emotions, e.g. 'anger' or 'neutrality', involves more universal acoustic cues. However, for some non-basic emotions, for instance, 'sarcasm' or 'happiness', perception is more culture-or-language specific. According to the confusion matrices of the naïve Dutch listeners and the advanced Dutch L2 learners of Chinese, we can infer that L1 influences L2 learners' perception of emotional prosody in the L2 to some extent. It further supports that perception of emotional prosody either in one's L1 or in one's L2 is partly universal and partly language-or-culture specific. Moreover, the third judgment study showed a strong in-group effect. It

confirms that emotion is generally better recognized between native listeners and native speakers from the same cultural group (e.g. Kilbride & Yarczower 1983, Markham & Wang 1996). However, the Chinese novice listeners in the third judgment study were not able to identify the emotional prosodies in an unknown language (Dutch) above chance level. Therefore, we can conclude that perception of emotional prosody in an unknown language is only universal to some extent – and that it is rather more cultural-or-language specific than universal.

From production point of view, L1 still plays a role in the production of emotional prosody, which means that L2 speakers of a target language might use L1-transfer to vocally express emotions in the target language. It seems that the advanced L2 speakers have already created their own approach (the hybrid system) to portray recognizable emotional prosody in the L2 after being learning the target language for several years. However, when L2 speakers were asked to produce emotional prosody in their L1, they expressed the emotional prosody much better than they did in the L2. It shows that production of emotional prosody is more language specific. However, for some basic emotions, such as 'anger', 'sadness' or 'neutrality', the acoustic correlates that L1 and L2 speakers used were very similar, implying that vocal production of emotion is universal to some extent. In summary, production of emotional prosody is more language-or-culture specific, although it is universal to some extent for some basic emotions. Moreover, different cultural groups have their own manner to express emotional prosody, which cannot be acquired by L2 speakers of that language along the increase of their proficiency of the target language.

## 8.3 General discussion

I carried out three judgment studies in this dissertation to answer the research questions. The eight research questions including the sub-questions of each research question listed in Chapter 1 have been carefully answered one by one. However, there are still ambiguities which need to be clarified in the future:

1) The functional view needs to be further tested

In the beginning of this dissertation, I introduced a functional view which claims that the prosodic space which languages may use is finite. Therefore, if a language uses vocal pitch for lexical purposes (i.e. for the marking of lexical tone), pitch will be used to a lesser extent for the expression of paralinguistic contrasts, such as signaling emotion. It could also be interpreted as that listeners of a tonal language might be less intent on (and in fact less experienced in) decoding the paralinguistic use of prosody than listeners of a non-tonal language. The advanced Dutch L2 learners of Chinese performed better than the native Chinese listeners when perceiving the emotions in Chinese. Moreover, novice Dutch listeners recognized the Chinese emotional prosodies as well as the native listeners did. These results altogether suggest that this functional view might be true, indicating that listeners of a non-tonal language are generally better than listeners of a tonal language at recognizing vocal emotions. However, this dissertation only investigated one tonal language (Mandarin Chinese) in terms of the production and the perception of emotional prosody. Therefore, it is difficult to know

at this stage whether the functional view is universal or not. It may be universal to some extent; for instance, it explained the unexpected results that the Dutch listeners outperformed the Chinese listeners or performed at least as well as the native listeners did. However, in order to get a complete picture of whether the functional view applies universally, future studies should investigate more linguistic groups. For instance, future studies can test how well listeners of a western non-tonal language (e.g. German, English or Spanish) perceive vocal emotion expressed in an Asian tonal language (Thai, Lao or Vietnamese) or in a sub-Saharan African language (Wolof, Koyra or Fulani) and vice versa.

2) 'L2 limits the expression of vocal emotion.' should be further investigated

The results of the first two judgment studies suggest that L2 limits the expression of vocal emotion. However, one may argue that it was not L2 which limited the expression of vocal emotion in the present study; actually, it might be the lexical tones which constrained the Dutch L2 speakers of Chinese to vocally produce emotions in the L2. This argument may be true, but it is not possible to purely separate lexical tone from emotional prosody in practice, since tone and (emotional) prosody both are phonetic suprasegments and they interact and entangle with each other in the daily communication. Therefore, it is suggested that future studies should involve some other L2 speaker groups whose L2 is a non-tonal language (e.g. Dutch L2 speakers of French) to further investigate whether it is the L2 or the lexical tone or both which limit the expression of vocal emotion in the L2.

3) More acoustic correlates need to be further studied

In this dissertation I only examined eight acoustic parameters, i.e. tempo, mean fundamental frequency, standard deviation of fundamental frequency, slope of the fundamental frequency, compactness, standard deviation of intensity, jitter and Harmonics to Noise Ratio in which tempo, F0, intensity and jitter were proved to be important variables in the production of vocal emotions in the previous studies. I also ran the automatic recognition of the six emotional prosodies to test whether the Linear Discriminant Analysis reflects the perception of emotional prosody by human listeners. The results showed that the LDA reflected human perception of vocal emotion to some extent. However, the human perception of the six emotional prosodies was still different from the perception of the LDA. It shows that some acoustic parameters used by LDA might not be used by the human listeners or vice versa. There may also be some other acoustic parameters contributing to the production and the perception of emotional prosody in general, which have been missed in this dissertation. Therefore, it is necessary for future studies to investigate more acoustic variables in the production of vocal emotion in speakers' (non)-native language; or adopt other methodologies to further study what factors influence the production of vocal emotion in an L2.

4) The hybrid system needs to be further tested

The results of the acoustic analysis showed a hybrid system developed by the L2 speakers of Chinese. This hybrid system needs to be further investigated, since the present study only studied one L2 speaker group, i.e. Dutch L2 speakers of Chinese. Therefore, it is not clear whether this hybrid solution is adopted universally by L2 speakers of other languages when they vocally produce emotions in the L2. Therefore, it is worth re-running the production experiment with some other L2 speaker groups, including both speakers of tonal and non-tonal language as a second language. For instance, one can further test how well English L2 speakers of French or English L2 speakers of Thai produce emotional prosody in French or Thai. The more L2 speaker groups are investigated, the clearer the picture is. However, I suspect that it is not easy for researchers to find L2 speakers of a tonal language whose native language is a non-tonal language to participate in a production experiment as subjects, e.g. English L2 speakers of Thai or Vietnamese, as these tonal languages are not commonly taught as foreign languages in the west.

# References


Albas, D. C., K. W. McCluskey & C. A. Albas (1976). Perception of the emotional content of speech: a comparison of two Canadian groups. *Journal of Cross-Cultural Psychology*, 7, 481–489.

Ang, J., R. Dhillon, E. Shriberg & A. Stolcke (2002). Prosody-based automatic detection of annoyance and frustration in human-computer dialog. *Proceedings of Interspeech, Denver*, 2037–2040.

Anolli, L., L. Wang, F. Mantovani & A. De Toni (2008). The voice of emotion in Chinese and Italian young adults. *Journal of Cross-cultural Psychology*, 39, 565–598.

Bachorowski, J. A. (1999). Vocal expression and perception of emotion. *Current Directions in Psychological Science*, 8, 53–57.

Bachorowski, J. A. & M. J. Owren (2008). Vocal expression of emotion. In M. Lewis, J. M. Haviland-Jones & L. F. Barrett (eds.) *Handbook of emotions*. New York: The Guilford Press, 196–210.

Banse, R. & K. R. Scherer (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70, 614–636.

Bänziger, T., D. Grandjean & K. R. Scherer (2009). Emotion recognition from expressions in face, voice, and body: the multimodal emotion recognition test (MERT). *American Psychological Association*, 9, 691–704.

Bent, T. & A. R. Bradlow (2003). The interlanguage speech intelligibility benefit. *Journal of the Acoustical Society of America*, 114, 1600–1610.

Berinstein, A. E. (1979). Stress. *Working Papers, University of California Los Angeles*, 47, 1–59.

Biehl, M., D. Matsumoto, P. Ekman & V. Hearn (1997). Matsumoto and Ekman's Japanese and Caucasian facial expression and emotion (JACFEE): reliability data and cross national differences. *Journal of Nonverbal Behavior*, 21, 3–21.

Biersack, S. & V. Kempe (2005). Tracing vocal expression of emotion along the speech chain: do listeners perceive what speakers feel? *Proceedings of ISCA Workshop on Plasticity in Speech Perception (PSP2005). London,* 211–214.

Blanc, J. & P. Dominey (2003). Identification of prosodic attitudes by a temporal recurrent network. *Cognitive Brain Research*, 17, 693–699.

Boersma, P. P. G. & D. J. M. Weenink (1996). *Praat, a system for doing phonetics by computer, version 3.4*. Report 132. Institute of Phonetic Sciences, University of Amsterdam (up-to-date version of the manual at http://www.fon.hum.uva.nl/praat/).

Booij, G. E. (1995). *The phonology of Dutch*. Oxford: Clarendon.

Borden, G. J., K. S. Harris & L. J. Raphael (1994). *Speech science primer: Physiology, acoustics and perception of speech* (3rd edition). Baltimore: Williams and Wilkins.

Castro, S. L. & C. Lima (2010). Recognizing emotions in spoken language: A validated set of Portuguese sentences and pseudosentences for research on emotional prosody. *Behavior Research Methods*, 42, 74–81.

Chang, F. M. (1985). Chinese culture and mental health. *International Journal of Psychiatry*, 3, 14–19.

Chao, Y. R. (1948). *Mandarin primer*. Cambridge: Harvard University Press.

Chen, A. J. (2005). *Universal and language-specific perception of paralinguistic intonational meaning*. LOT Dissertation Series 102. Utrecht: LOT.

Cheng, C. C. (1997). Measuring relationship among dialects: DOC and related resources. *Computational Linguistics & Chinese Language Processing*, 2, 41–72.

Chuang, C. K., S. Hiki, T. Sone & T. Nimura (1972). The acoustical features and perceptual cues of the four tones of standard colloquial Chinese. *Proceedings of the 7th International Congress on Acoustics (volume 3)*. Budapest: Akademial Kiado, 297–300.

Colombetti, G. (2009). From Affect Programs to Dynamical Discrete Emotions. *Philosophical Psychology*, 22, 407–425.

Comrie, B., M. S. Dryer., G. David & M. Haspelmath (2005). *The world atlas of language structures*. Oxford: Oxford University Press.

Cornew, L., L. Carver & T. Love (2010). There's more to emotion than meets the eye: A processing bias for neutral content in the domain of emotional prosody. *Cognition and Emotion*, 24, 1133–1152.

Cowie, R., E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey & M. Schröder (2000). Feeltrace: An instrument for recording perceived emotion in real time. In R. Cowie, E. Douglas-Cowie & M. Schröder (eds.) *A conceptual framework for research. Proceedings of the ISCA Workshop on Speech and Emotion, Belfast*, 19–24.

Cruttenden, A. (1986). *Intonation*. Cambridge: Cambridge University Press.

Darwin, C. (1998). *The expression of the emotions in man and animals*. New York: Oxford University Press. (original work published 1872).

Davis, H. (1976). Principles of electric response audiometry. *Annals of Otology, Rhinology and Laryngology*, 85, suppl. 28.

Deller, J. R., J. G. Proakis & J. H. L. Hansen (1993). *Discrete-time processing of speech signals*. New York: Macmillan.

Dennis, D. G. (1982). Ethnicity and the capacity to interpret nonverbal communication. (Doctoral dissertation, Yeshiva University). *Dissertation Abstracts International*, 43, 1301.

De Pijper, J. R. (1983). *Modeling British English intonation*. Dordrecht: Foris.

Dreher, J. J. & P. C. Lee (1966). *Instrumental investigation of single and paired Mandarin tonemes*. Research Communication No. 13) Huntington Beach, CA: Advanced Research Laboratory, Douglas Aircraft Company.

Dreher, J. J., E. L. Young & P. C. Lee (1969). *Mandarin triplet contours*. Research Communication No. 107. Huntington Beach, CA: Advanced Research Laboratory, Douglas Aircraft Company.

Dromey C., J. Silveira & P. Sandor (2004). Recognition of affective prosody by speakers of English as a first or foreign language. *Speech Communication*, 47, 351–359.

Duanmu, S. (2006). Chinese (Mandarin): phonology. In K. Brown (2nd edition) *Encyclopedia of language and linguistic*s. Oxford: Elsevier Publishing House.

Duanmu, S. (2007a). Stress, information and language typology. *Yuyan Kexue*, 6, 3–16.

Duanmu, S. (2007b). *The phonology of standard Chinese*. Oxford: Oxford University Press.

Ekman, P. (1972). Universals and cultural differences in facial expressions of emotion. In J. Cole (ed.) *Nebraska symposium on motivation, 1971*. Lincoln, NE: University of Nebraska Press (volume 19), 207–282.

Ekman, P. & W. Friesen (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17 124–129.

Elfenbein, H. A. & N. Ambaby (2002). On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological Bulletin*, 128, 203–235.

Fichten, C. S., V. Taglakis, D. Judd, J. Wright & R. Amsel (1992). Verbal and nonverbal communication cues in daily conversations and dating. *The Journal of Social Psychology*, 132, 751–769.

Flege, J. E. (1997). The role of phonetic category formation in second-language speech learning. In J. Leather & A. James (eds.) *New Sounds 97. Proceedings of the Third International Symposium on the Acquisition of Second-Language Speech, Amsterdam*, 79–88.

Flege, J. E., N. Takagi & V. Mann (1995). Japanese adults can learn to produce English /r/ and /l/ accurately. *Language and Speech*, 38, 25–55.

Frick, R.W. (1985). Communicating emotion: the role of prosodic features. *Psychological Bulletin*, 97, 412–429.

Frijda, N. (2000). The psychologist's view. In M. Lewis, J. M. Haviland-Jones & L. F. Barrett (eds.) *Handbook of emotions*. New York: Guilford Press, 59–74.

Gandour, J. T. (1983). Tone perception in Far Eastern languages. *Journal of Phonetics*, 11, 149–175.

Gitter, A. G., H. Black & D. I. Mostofsky (1972). Race and sex in the communication of emotion. *Journal of Social Psychology,* 88*,* 273–276.

Gooskens, C. & R. van Bezooijen (2006). Mutual comprehensibility of written Afrikaans and Dutch: Symmetrical or asymmetrical? *Literary and Linguistic Computing*, 23, 543–557.

Gorelick, P. & E. Ross (1987). The aprosodias: Further functional-anatomical evidence for the organization of affective language in the right hemisphere. *Journal of Neurology, Neurosurgery and Psychiatry*, 50, 553–560.

Graham, C. R., A. W. Hamblin & S. Feldstein (2001). Recognition of emotion in English voices by speakers of Japanese, Spanish and English. *International Review of Applied Linguistics in Language Teaching*, 39, 19–37.

Grandjean, D., T. Bänziger & K. R. Scherer (2006). Intonation as an interface between language and affect. In S. Anders, G. Ende, M. Junghofer, J. Kissler & D. Wildgruber (eds.) *Understanding emotion*. Amsterdam: Elsevier, 235–248.

Haidt, J. & D. Keltner (1999). Culture and facial expression: open-ended methods find more expressions and a gradient of recognition. *Cognition and Emotion*, 13, 225–266.

Hart, J. 't, R. Collier & A. Cohen (1990). *A perceptual study of intonation*. Cambridge: Cambridge University Press.

Hess, U. & P. Thibault (2009). Darwin and emotion expression. *American Psychologist*, 64, 120–128.

Howie, J. M. (1970). The vowels and tones of Mandarin Chinese: Acoustical measurements and experiments. Ph.D. dissertation, Indiana University.

Howie, J. M. (1976). *Acoustical studies of Mandarin vowels and tones*. Cambridge: Cambridge University Press.

Huttar, G. L. (1968). Relations between prosodic variables and emotions in normal American English utterances. *Journal of Speech and Hearing Research*, 11, 481–487.

Johnson, W. F., R. N. Emde, K. R. Scherer & M. D. Klinnert (1986). Recognition of emotion from vocal cues. *Archives of General Psychiatry*, 43, 280–283.

Johnstone, T. & K. R. Scherer (2000). Vocal communication of emotion. In M. Lewis, J. M. Haviland-Jones & L. F. Barrett (eds.) *Handbook of emotions*. New York: The Guilford Press, 220–235.

Jongman, A., Y. Wang, C. Moore & J. A. Sereno (2006). Perception and production of Mandarin tones. In P. Li, H. T. Li, E. Bates & O. J. L. Tzeng (eds.) *Handbook of Chinese Psycholinguistics*. Cambridge, UK: Cambridge University Press, 209–216.

Juslin, P. N. & P. Laukka (2001). Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. *Emotion*, 1, 381–412.

Juslin, P. N. & P. Laukka (2003). Communication of emotions in vocal expression and music performance: different channels, same code? *Psychological Bulletin*, 129, 770–814.

Kager, R. W. J. (1989). *A metrical theory of stress and destressing in English and Dutch*. Foris: Dordrecht.

Kilbride, J. E. & M. Yarczower (1983). Ethnic bias in the recognition of facial expressions. *Journal of Nonverbal Behavior*, 8, 27–41.

Klecka, W. R. (1980). *Discriminant analysis*. Quantitative applications in the social sciences. Beverly Hills, CA: Sage Publications.

Klineberg, O. (1938). Emotional expression in Chinese literature. *Journal of Abnormal and Social Psychology*, 33*,* 517–520.

Kraayeveld, J. (1997). Idiosyncrasy in prosody: Speaker and speaker group identification in Dutch using melodic and temporal information. PhD dissertation, Catholic University Nijmegen.

Ladd, D. R., K. E. A. Silverman, F. Tolkmitt, G. Bergmann & K. R. Scherer (1985). Evidence for the independence of intonation contour type, voice, quality, and F0 range in signalling speaker affect. *Journal of the Acoustical Society of America,* 78, 435–444.

Langeweg, S. J. (1988). The stress system of Dutch. PhD dissertation, Leiden University.

Leather, J. (1990). Perceptual and productive learning of Chinese lexical tone by Dutch and English speakers. In J. Leather & A. James (eds.) *New Sounds 90. Proceedings of the Amsterdam Symposium on the Acquisition of Second Language Speech, Amsterdam*, 305–341.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge MA: MIT Press.

Leinonen, L., T. Hiltunen, I. Linnankoski & M-L. Laakso (1997). Expression of emotional-motivational connotations with a one-word utterance. *Journal of the Acoustic Society of America*, 102, 1853–1863.

Li, A., P. F. Shao & J. W. Dang (2009). Cross-cultural and multi-modal investigation of emotion expression. *Journal of Tsinghua University*, 49, 1393–1401.

Li, R. (1987, a-2). Chinese dialects in China. *The Language Atlas of China* (English version).

Lieberman, P. & B. Michaels (1962). Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech. *Journal of the Acoustical Society of America*, 34, 922–927.

Liu, F. (1924). *Szu Sheng Shih Yen Lu*. Shanghai: Ch'un Yi.

Liu, P. & M. D. Pell (2012). Recognizing vocal emotions in Mandarin Chinese: A validated database of Chinese vocal emotional stimuli. *Behavior Research Methods*, 44, 1042–1051.

Markham, R. & L. Wang (1996). Recognition of emotion by Chinese and Australian children. *Journal of Cross-cultural Psychology*, 27, 616–643.

McCluskey, K. W., D. C. Albas, R. R. Niemi, C. Cuevas & C. A. Ferrer (1975). Cross-cultural differences in the perception of the emotional content of speech. A study of the development of sensitivity in Canadian and Mexican children. *Developmental Psychology*, 11, 551–555.

Mitchell, R. L. C. & E. D. Ross (2013). Attitudinal prosody: What we know and directions for future study. *Neuroscience and Biobehavioral Reviews*, 37, 471–479.

Monrad-Kohn, G. H. (1947). The prosodic quality of speech and its disorders (a brief survey from a neurologist's point of view). *Acta Psychiatrica et Neurologica Scandinavica*, 22, 255–269.

Monrad-Kohn, G. H. (1963). The third element of speech: Prosody and its disorders. In L. Halpern (ed.) *Problems of dynamic neurology*. Jerusalem: Department of Nervous Diseases of the Rothschild Hadassah, 101–118.

Moore, C. B. & A. Jongman (1997). Speaker normalization in the perception of Mandarin Chinese tones. *Journal of the Acoustical Society of America*, 102, 1864–1877.

Mozziconacci, S. J. L. (2001). Modeling emotion and attitude in speech by means of perceptually based parameter values. *User Modeling and User-Adapted Interaction*, 11, 297–326.

Nederlandse Taalunie (2005). http://taalunieversum.org/inhoud/feiten-en-cijfers.

Nooteboom, S. G. (1997). The prosody of speech: Melody and rhythm. In W. J. Hardcastle & J. Laver (eds.) *The handbook of phonetic sciences*. Oxford: Blackwell, 640–673.

Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica*, 41, 1–16.

Ohman, A., A. Flykt & F. Esteves (2001). Emotion drives attention: Detecting the snake in the grass. *Journal of Experimental Psychology: General*, 130, 466–478.

Owren, M. J. & D. Rendall (1997). An affect-conditioning model of nonhuman primate signaling. In D. H. Owings, M. D. Beecher & N. S. Thompson (eds.) *Perspectives in Ethology*, *Volume 12: Communication*. New York: Plenum Press, 299–346.

Owren, M. J. & D. Rendall (2001). Sound on the rebound: Bringing form and function back to the forefront in understanding nonhuman primate vocal signaling. *Evolutionary Anthropology*, 10, 58–71.

Owren, M. J., D. Rendall & J.-A. Bachorowski (2003). Nonlinguistic vocal communication. In D. Maestripieri (ed.) *Primate psychology*. Cambridge, MA: Harvard University Press, 359–394.

Pakosz, M. (1983). Attitudinal judgments in intonation: Ssome evidence for a theory. *Journal of Psycholinguistic Research*, 12, 311–326.

Pell, M. D. (2001). Influence of emotion and focus location on prosody in matched statements and questions. *Journal of the Acoustical Society of America*, 109, 1668–1680.

Pell, M. D. (2006). Judging emotion and attitudes from prosody following brain damage. *Progress in Brain Research*, 156, 303–317.

Pell, M. D., L. Monetta, S. Paulmann & S. A. Kotz (2009). Recognizing emotions in a foreign language. *Journal of Nonverbal Behavior*, 33, 107–120.

Pell, M. D. & V. Skorup (2008). Implicit processing of emotional prosody in a foreign versus native language. *Speech Communication*, 6, 519–530.

Pike, K. (1948). *Tone languages: a technique for determining the number and type of pitch contrasts in a language, with studies in tonemic substitution and fusion*. University of Michigan Publications in Linguistics, 4. Ann Arbor: University of Michigan Press.

Pinto, N. B. & I. R. Titze (1990). Unification of perturbation measures in speech signals. *Journal of the Acoustical Society of America*, 87, 1278–1289.

Potisuk, S., J. Gandour & M. Harper (1996). Acoustic correlates of stress in Thai. *Phonetica*, 53, 200–220.

Remijsen, B. (2002a). *Word-prosodic systems of Raja Ampat languages*. LOT Dissertation Series 49. Utrecht: LOT.

Remijsen, B. (2002b). Lexically contrastive stress accent and lexical tone in Ma'ya. In C. Gussehoven & N. Warner (eds.) *Laboratory phonology 7* Berlin/New York: Mouton de Gruyter, 585–614.

Rendall, D. & M. J. Owren (2002). Animal vocal communication: say what? In M. Bekoff, C. Allen & G. Burghardt (eds.) *The cognitive animal*. Cambridge, MA: MIT Press, 307–314.

Rosenberg, E. L. & P. Ekman (1995). Conceptual and methodological issues in the judgment of facial expressions of emotion. *Motivation & Emotion,* 19, 111–138.

Ross, E. D., J. A. Edmondson & G. B. Seibert (1986). The effect of affect on various acoustic measures of prosody in tone and non-tone languages: A comparison based on computer analysis of voice. *Journal of Phonetics*, 14, 283–302.

Ross, E. D. (2000). Affective prosody and the aprosodias. In M.-M. Mesulam (ed.) *Principles of behavioral and cognitive neurology*. New York: Oxford University Press, 316–331.

Rumjancev, M. K. (1972). Ton i intonacija v sovremennom kitajskom Jazyke [Tone and intonation in Modern Chinese] (Izdatel'stvo Moskovskogo Universiteta, Moscow). Reviewed by A. V. Lyovin (1978). *Journal of Chinese Linguistics*, 6, 120–168.

Scherer, K. R. (1979). Nonlinguistic vocal indicators of emotion and psychopathology. In C. E. Izard (ed.) *Emotions in personality and psychopathology*. New York: Plenum, 493–529.

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, 99, 143–165.

Scherer, K. R. (1989). Vocal correlates of emotional arousal and affective disturbance. In H. Wagner & A. Manstead (eds.) *Handbook of psychophysiology: Emotion and social behavior*. London: Wiley, 165–197.

Scherer, K. R. (2000). Emotions as episodes of subsystem synchronization driven by nonlinear appraisal processes. In M. D. Lewis & I. Granic (eds.) *Emotion, development, and self-organization*. Cambridge: Cambridge University Press, 70–99.

Scherer, K. R. (2003). Vocal communication of emotion: a review of research paradigms. *Speech Communication*, 40, 227–256.

Scherer, K. R., R. Banse & H. G. Wallbott (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology*, 32, 76–92.

Scherer, K. R., R. Banse, H. G. Wallbott & T. Goldbeck (1991). Vocal cues in emotion encoding and decoding. *Motivation and Emotion*, 15, 123–148.

Scherer, K. R., D. R. Ladd & K. E. A. Silverman (1984). Vocal cues to speaker affect: Testing two models. *Journal of the Acoustical Society of America*, 76, 1346–1356.

Scherer, K. R., H. G. Wallbott & A. B. Summerfield (1986). *Experiencing emotion: A cross-cultural study*. Cambridge: Cambridge University Press.

Schmitt, J. J., W. Hartje & K. Willmes (1997). Hemispheric asymmetry in the recognition of emotional attitude conveyed by facial expression, prosody and propositional speech. *Cortex*, 33, 65–81.

Schröder, M. (2000). Experimental study of affect bursts. In R. Cowie, E. Douglas-Cowie & M. Schröder (eds.) *A conceptual framework for research. Proceedings of the ISCA Workshop on Speech and Emotion, Belfast*, 132–137.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics*, 10, 209–241.

Shochi, T., G. Gagnié, A. Rilliard, D. Erickson & V. Aubergé (2010). Learning effect of prosodic social affects for Japanese learners of French language. Paper presented at the 5th International Conference on Speech Prosody, 11-14 May, Chicago.

Sobin, C. & M. Alpert (1999). Emotion in speech: The acoustic attributes of fear, anger, sadness, and joy. *Journal of Psycholinguistic Research*, 28, 347–365.

Spackman, M. P., B. L. Brown & S. Otto (2009). Do emotions have distinct vocal profiles: a study of idiographic patterns of expression. *Cognition and Emotion*, 23, 1565–1588.

Speech Therapy Information and Resources (2008). http://www.speech-therapy-information-and-resources.com/acoustic-measures-norms.html.

Stagray, J. & D. Downs (1993). Differential sensitivity for frequency among speakers of a tone and a nontone language. *Journal of Chinese Linguistics*, 21, 143–163.

Standke, R. (1992). *Methoden der digitalen Sprachverarbeitung in der vokalen Kommunikations-forschung [Methods of digital speech processing in research on vocal communication]*. Frankfurt: Peter Lang.

Streeter, L. A., N. H. Macdonald, W. Apple, R. M. Krauss & K. M. Galotti (1983). Acoustic and perceptual indicators of emotional stress. *Journal of the Acoustical Society of America*, 73, 1354–1360.

Tang, C. (2009). *Mutual intelligibility of Chinese dialects: An experimental approach.* LOT Dissertation Series 228. Utrecht: LOT.

Tang, C. & V. J. Van Heuven (2009). Mutual intelligibility of Chinese dialects experimentally tested. *Lingua*, 119, 709–732.

Tartter, V. C. & D. Braun (1994). Hearing smiles and frowns in normal and whisper registers. *Journal of the Acoustical Society of America*, 96, 2101–2107.

Thompson, W. F. & L-L. Balkwill (2006). Decoding speech prosody in five languages. *Semiotica*, 158, 407–424.

Tickle, A. (2000). English and Japanese speaker's emotion vocalizations and recognition: a comparison highlighting vowel quality. In R. Cowie, E. Douglas-Cowie & M. Schröder (eds.) *A conceptual framework for research. Proceedings of the ISCA Workshop on Speech and Emotion, Belfast*, 104–109.

Tompkins, C. A. & C. A. Mateer (1985). Right hemisphere appreciation of prosodic and linguistic indications of implicit attitude. *Brain and Language*, 24, 185–203.

Tooby, J. & L. Cosmides (1990). The past explains the present: emotional adaptations and the structure of ancestral environments. *Ethology and Sociobiology*, 11, 375–424.

Van Bezooijen, R. (1984). *The characteristics and recognizability of vocal expression of emotions*. Dordrecht: Foris.

Van Heuven, V. J. (1994). What is the smallest prosodic domain? In P. Keating (ed.) *Papers in Laboratory Phonolgy III: phonological structure and phonetic form*. London: Cambridge University Press, 76–98.

Van Heuven, V. J. & H. Van de Velde (2010). De uitspraak van het hedendaags Nederlands in de Lage Landen [The pronunciation of present-day Dutch in the Low Countries]. In J. Fenoulhet & J. Renkema (eds.) *Internationale neerlandistiek: Een vak in beweging* (Lage Landen Studies, 1). Gent: Academia Press, 183–209.

Van Santen, J. P. H., E. T. Prud'hommeaux & L. M. Black (2009). Automated assessment of prosody production. *Speech Communication*, 51, 1082–1097.

Van Zanten, E. A. & R. W. N. Goedemans (2007). A functional typology of Austronesian and Papuan stress systems. In V. J. van Heuven & E. A. van Zanten (eds.) *Prosody in Indonesian languages*. LOT Occasional Series, 7. Utrecht: LOT, 63–88.

Wallbott, H. G. & K. R. Scherer (1986). Cues and channels in emotion recognition. *Journal of Personality and Social Psychology*, 51, 690–699.

Wang, W. S.-Y. & K.-P. Li (1967). Tone 3 in Pekinese. *Journal of Speech and Hearing Research*, 10, 629–636.

Wang, Y., M. Spence, A. Jongman & J. Sereno (1999). Training American listeners to perceive Mandarin tones. *Journal of the Acoustical Society of America*, 106, 3649–3658.

Wu, D. & W.-C. Tseng (1985). *Introduction: The characteristics of Chinese culture*. Orlando, FL: Academic Press.

Wu, Z. J. (1986). *The spectrographic album of mono-syllables of standard Chinese*. Beijing: Social Science Press.

Xing, F. Y. (1999). *汉语语法特点面面观 [Diverse aspects of Chinese syntax characteristics]*. Beijing: Beijing Language and Culture University Press.

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of F0 contours. *Journal of Phonetics*, 27, 55–105.

Yip, M. (2006). Tone. In P. de Lacy (ed.) *The Cambridge handbook of phonology*. Cambridge: Cambridge University Press, 229–252.

You, M. Y., C. Chen & J. J. Bu (2005). CHAD: A Chinese affective database. *Proceedings of the 1st International Conference on Affective Computing and Intelligent Interaction. Beijing*, 542–549.

Zhang, S., P.-C. Ching & F.-R. Kong (2006). Acoustic analysis of emotional speech in Mandarin Chinese. *Proceedings of the International symposium on Chinese Spoken Language Processing*. Singapore: Kent Ridge, 57–66.

Zhu, Y. (2013a). Which is the best listener group? Perception of Chinese emotional prosody by Chinese natives, naïve Dutch listeners and Dutch L2 learners of Chinese, *Dutch Journal of Applied Linguistics*, 2, 170–183.

Zhu, Y. (2013b). Perception of Chinese emotional prosody produced by Dutch learners and native speakers of Chinese. *Chinese as a Second Language Research* (accepted).

# Summary

Since Darwin published his book *The Expression of the Emotions in Man and Animals* in 1872, there has been an increasing number of studies on perception and production of emotions. Earlier studies were mainly conducted in the fields of psychology, physiology, biology; later they were extended into other areas, such as sociology, linguistics, pathology, computer science, neuroscience, musicology and second language acquisition. In addition, there have always been studies on perception or production of emotion cross-culturally or cross-linguistically. It is claimed by some researchers that perception of emotion is universal. However, other researchers believe that it is partly universal and partly cultural-or-language specific. Previous studies showed that emotion is generally better recognized when expressed by a speaker of the same cultural group as the listeners (in-group advantage). These studies also indicated that automatic recognition of human-produced emotions can reveal some of the acoustic cues that humans use to perceive and produce emotions. Although previous findings of perception and production of emotion are abundant, previous studies do not give us a clear picture of how well listeners of a non-tonal language can perceive emotions produced in a tonal language (especially through the audio channel). Neither do earlier studies give us a clear view of how well non-native speakers of a language can vocally produce emotion in the second or foreign language (L2) compared to native speakers, especially when the L2 is a tonal language but the L2 speakers' L1 is not. It is also not clear whether a speaker can vocally produce emotions in his L2 as well as he does in his native language (L1).

Therefore, the first aim of this dissertation was to investigate experimentally how well native (Mandarin) and non-native (Dutch) listeners of a tonal language (Mandarin) perceive vocal emotions portrayed in the tonal language. Non-native listeners included both naïve listeners and advanced L2 learners of Mandarin who shared the same L1 (Dutch) as naïve listener group. Secondly, I investigated whether Dutch L2 speakers of Mandarin are able to vocally produce emotions in the L2 as well as they do in their non-tonal L1; at the same time, I studied how well naïve native listeners and advanced learners of Mandarin perceive vocal emotion expressed by L2 speakers of the tonal language. An acoustic analysis was later conducted to analyze vocal correlates that speakers and listeners use in the production and perception of the vocal emotions. Finally, I investigated whether the in-group advantage reported by other researchers (claiming that listeners generally better recognize emotional prosody produced in their L1 than in an unknown language) is universal.

From a theoretical point of view, this dissertation aims to test a functional hypothesis, claiming that the prosodic space which languages may use, is finite. Therefore, if a language uses duration to mark a segmental contrast between long and short vowels, the duration parameter will not play a role (or a less important role) in the marking of stress – which in other languages depends rather heavily on duration cues (Berinstein 1979, Potisuk et al. 1999, Remijsen 2002a, b). By the same token, if a language such as

Mandarin uses pitch for lexical purposes (i.e. lexical tone), less room will be left for the signaling through pitch of paralinguistic contrasts, such as the expression of emotion. As a consequence of this I predict that native listeners of Mandarin will have limited exposure to clear exemplars of prosodically expressed affect. More generally, I predict that native listeners of a tonal language will be less intent on (and in fact less experienced in) decoding paralinguistic use of prosody than listeners of a non-tonal language.

In the introductory **Chapter 1** I summarize what has been done by previous studies, and then identify some unsolved issues. Background knowledge on tonal and non-tonal language, specifically Mandarin Chinese and Dutch, tone and (emotional) prosody, and the acoustic aspects of emotional prosody is provided in this chapter, too. I then itemize my research questions and derive specific predictions from the functional view mentioned above. I also motivate the choice of six emotional prosodies ('neutral', 'happy', 'angry', 'surprised', 'sad' and 'sarcastic') for the present study and propose feasible research methods needed to answer my the research questions. The end of this chapter provides an outline of the dissertation.

**Chapter 2** reviews previous studies in more detail. There are many on the perception of emotional prosody within the same cultural group; however, research on cross-cultural perception of vocal emotion is relatively scarce. There are even fewer cross-linguistic studies on the production of vocal emotion. And there seem to be no earlier studies which directly compare the production of emotional prosody in a speaker's L1 and L2. Therefore, the production part in this dissertation is pioneering.

Many researchers have claimed that listeners can recognize emotional prosody above chance level both within-culturally and cross-culturally. However, emotion is generally better identified between speakers and listeners who are from the same cultural groups. The accuracy of recognition decreases as the cultural distance between two cultural groups is bigger. Previous studies also found that successful communication of vocal emotion depends on both the speaker and the listener, although the role of the listener seems to be more important than the role of the speaker. Some researchers indicate that production of emotional prosody cross-culturally may be universal to some extent; however, production of some emotions is cultural-specific.

In this chapter, I also provide information on the methodology used. There are three judgment studies in this dissertation. The first judgment study includes one perception experiment (Exp. 1), in which native Chinese listeners, naïve Dutch listeners and advanced Dutch learners of Chinese perceived and identified the six Chinese emotional prosodies (see above) portrayed by native Chinese speakers. This experiment aimed to find an answer to research question (i) *How well can native Chinese, Dutch naïve listeners and advanced Dutch learners of Chinese perceive Chinese emotional prosodies vocally portrayed by Chinese native speakers? What will be the confusion patterns of the three listener groups?* The second judgment study included two perception experiments: in the first the same three listener groups listened to the same six Chinese emotional prosodies but produced by Dutch L2 speakers of Chinese (Exp. 2A). This experiment was designed to find answers to research question (ii) *How well can native Chinese, Dutch naïve listeners and advanced Dutch learners of Chinese perceive Chinese emotional prosodies vocally portrayed by Dutch L2 speakers of*

*Chinese? What will be the confusion patterns of the three listener groups?* In the second perception experiment (Exp. 2B), Dutch native listeners listened to the same six emotional prosodies portrayed in their native language (Dutch) by the same Dutch L2 speakers of Chinese. This was to test how well the same Dutch L2 speakers of Chinese produce the emotional prosodies in their L1. The results of this perception experiment were compared with the results obtained in the first perception experiment of the second judgment study to answer research questions (iii) *Can Dutch L2 speakers of Chinese produce emotional prosodies in the L2 as well as they do in the L1- Dutch? What will be the similarities and differences between these two types of production?* and (iv) *Does L2 limit the expression of emotional prosody, especially when the native language of L2 speakers of the tonal language is a non-tonal language?* Research question (v) *Is the functional view true, predicting that listeners of a tonal language might be less intent on (and in fact less experienced in) decoding the paralinguistic use of prosody than listeners of a non-tonal language?* was answered after the first and the second judgment study, questioning whether the functional view is true. An acoustic analysis was made of the stimuli used in the two judgment studies. The results answered the research question (vi) *What acoustic parameters contribute to differentiate between emotional prosodies in general? What acoustic correlates do speakers and listeners use to produce and perceive the vocal emotions in their L1 and in an L2? Do Dutch L2 speakers of Chinese use L1-transfer to produce emotional prosody in Chinese? To what extent automatic recognition reflects the perception of the emotional prosodies by the human listeners?* The third judgment study was conducted in a reciprocal way. It included two perception experiments in which Chinese and Dutch novice listeners perceived the six emotions vocally portrayed in their L1 and in the other language (Exp. 3). This experiment was designed to test whether the in-group advantage claimed by other researchers is universal, which was the research question (vii) *Is the in-group advantage universal, claiming that listeners are better in recognizing emotional prosody produced in their native language than in their L2 or an unknown language? Moreover, is the perception of vocal emotion cross-culturally symmetrical between Chinese and Dutch listeners, i.e., will Dutch and Mandarin listeners have similar abilities of identifying emotional prosody expressed in the other language?* The three judgment studies altogether answered the research question (viii) *Are perception and production of emotional prosody universal? Or are they more language-specific and culture specific?*

**Chapter 3** reports the results of the first judgment study (Exp. 1), which was used as a baseline for the later studies. Twenty Chinese native listeners, 20 naïve Dutch listeners and 20 advanced Dutch L2 learners of Chinese recognized the six Chinese emotional prosodies (neutrality, happiness, anger, surprise, sadness and sarcasm). The results show that advanced Dutch L2 learners of Chinese recognized Chinese emotional prosody significantly (54 % correct) better than Chinese native listeners (46 % correct) and Dutch naïve listeners (46 % correct). The results also indicate that naïve non-native (Dutch) listeners could recognize emotions in an unknown language (Mandarin) as well as the natives did. Chinese native listeners did not show an in-group advantage for identifying emotions in Chinese more accurately and confidently. 'Neutrality' was the easiest emotion for all the three listener groups to identify and 'anger' was recognized equally well by all the listener groups. The prediction made in the beginning of the study is confirmed: listeners of a tonal language will be less proficient in the para-linguistic use of prosody than listeners of a non-tonal language. The results in this chapter provide the baseline for Chapter 4.

**Chapter 4** investigates the differences between perception of six Chinese emotional prosodies (neutrality, happiness, anger, surprise, sadness and sarcasm) produced by Dutch L2 speakers of Chinese and those encoded by native Chinese speakers (control group). This chapter compares the results of Exp. 1 and Exp. 2A. Twenty Chinese native listeners, 20 naïve non-native listeners (Dutch) and 20 advanced Dutch L2 learners of Chinese listened to the Chinese emotional prosodies expressed by both L1 and L2 speakers of Chinese. The results show that the three listener groups recognized emotional prosodies encoded by Chinese natives significantly better (49% correct) than those produced by L2 speakers of Chinese (39% correct). Also, the naïve non-native listeners could recognize the emotions in the unknown language (46% correct) as well as the natives did. In terms of perceiving L2-produced Chinese emotional prosody, although the advanced Dutch L2 learners of Chinese performed slightly better than the other two listener groups, there was no significant effect among the three listener groups in terms of identifying L2-produced emotional prosody. The functional view is once again confirmed, which claims that listeners of a tonal language will be less proficient in the paralinguistic use of prosody than listeners of a non-tonal language. Therefore, in some cases at least, linguistic use of a particular acoustic feature in spoken language limits its use for the communication of emotion.

**Chapter five** gives the complete picture of how Dutch L2 speakers of Chinese produced the six emotional prosodies in their L2 (Chinese) and how they expressed the same emotional prosodies in their native language (Dutch). This chapter reports the results of the first and the second judgment study together; it was written from the production point of view. The results show that emotional prosodies produced by Dutch L2 speakers of Chinese were overall less recognizable ((39% correct by Chinese listeners) in their L2 than those encoded by Chinese natives (46% correct). Dutch L2 speakers of Chinese were better at vocally producing emotions in their L1 (57% correct by Dutch listeners). The prediction made in the beginning of the chapter is confirmed: speaking in an L2 limits the speaker's communication of emotion. The results also show that the naïve Dutch listeners were able to recognize the emotions in the unknown language as well as the natives Mandarin listeners did. Moreover, naïve Dutch listeners showed an in-group advantage: they identified emotions in Dutch more accurately than they did in Chinese.

**Chapter 6** presents an acoustic analysis of three types of production of emotional prosody: L1 Mandarin, L2 Mandarin and L1 Dutch (the latter two were produced by the same individuals). Eight acoustic correlates were examined: tempo, mean fundamental frequency (pitch or F0), the standard deviation of the pitch (SD_F0), rate of change of the F0 (slope_F0), compactness of the spectral energy distribution, the standard deviation of the intensity (SD_int), jitter (cycle-to-cycle variation of the glottal pulses) and HNR (harmonics-to-noise ratio in the vocal signal). I also performed an automatic recognition of the six emotional prosodies portrayed by the three speaker groups, using Linear Discriminant Analysis (LDA). The acoustic analysis shows that fundamental frequency, including mean F0, SD_F0 and slope_F0, is an influential variable in the production of vocal emotions by the three groups of speakers. This finding confirms the study of Scherer (1996), who claimed that F0 plays a crucial role in the production of emotional prosody. Jitter and standard deviation of the intensity did not contribute much to differentiating between emotions in the present study. 'Tempo'

and 'compactness' were only sensitive to Mandarin L1 speakers. Slope of the F0 indicates that Chinese uses rising intonation to express surprise, which confirms previous studies, claiming that many tonal languages use rising intonation to express surprise (Yip 2006). Moreover, HNR can clearly distinguish 'sad' from 'neutral' in L2 Mandarin and L1 Dutch (produced by the same individuals), but not in L1 Mandarin. In summary, fundamental frequency is a very influential variable in the production and perception of vocal emotion in general. Other parameters studied in this chapter also contribute to differentiating between emotional prosodies, but they are more emotion-specific or speaker-type specific.

The results of the LDA show that the human-produced emotional prosody can be automatically recognized from the acoustic measures well above chance level (50% overall correct). There was significant correlation between confusions obtained by the automatic recognition and by the human listeners in the present study, so that automatic recognition reflects the perception of emotional prosody by human listeners to some extent. However, there may also be some other acoustic parameters contributing to the production and the perception of emotional prosody in general, which have been missed in this dissertation.

The acoustic analysis indicates that Dutch L2 speakers use some acoustic parameters in the production of emotional prosody in the L2 (Chinese) the same way they do in their L1 (Dutch), e.g. SD_F0 and SD_Int. Therefore, we may conclude that L1-transfer is a strategy for L2 speakers to vocally produce emotions in the L2. However, this strategy may not work for all the emotions, e.g. not for 'surprise' and 'sarcasm'. Moreover, the results suggest that the L2 speakers did not automatically acquire the native approach to vocally produce emotional prosody in the target language during their L2 learning process. It seems that these advanced L2 speakers of Chinese have developed a hybrid system of producing emotional prosody in the L2. The hybrid system approximates the Chinese native manner of portraying vocal emotion to some extent (the way it involves tempo, mean F0, slope of the F0, compactness and jitter), but exploits the variability in F0 and intensity that the L2 speakers use to produce emotional prosody in their L1.

**Chapter 7** investigates the perception of emotional prosody by native and novice listeners in a reciprocal way (Exp. 3). Twenty Chinese and 20 Dutch native listeners who do not have any knowledge of Dutch and Chinese, identified the emotional prosodies in these two languages. The results showed that novice Dutch listeners (46% correct) could recognize emotional prosody in the unknown language (Chinese) as well as natives did (46% correct); and they performed significantly better in identifying emotional prosody expressed in their native language (Dutch, 57% correct). In contrast to this, Chinese novice listeners were only able to recognize emotional prosody in their L1 reasonably well (46% correct) but failed to identify vocal emotion in the unknown language (Dutch, 15% correct) above chance level. This finding confirms the existence of the in-group advantage found by other researchers, claiming that listeners generally better recognize emotional prosody produced in their L1 than in an unknown language. Moreover, the results suggest that perception of vocal emotion cross-culturally is not symmetrical, so that some cultural group might be generally better than some other cultural group at perceiving emotional prosody. This, again, lends credibility to the

functional view which predicts that listeners of a tonal language will generally be less proficient in the perception of vocal emotion than listeners of a non-tonal language.

**Chapter 8** reviews the main findings of this dissertation, and uses these to answer the research questions asked in Chapter 1. The chapter is concluded by a discussion of aspects that can be improved in the future.

# Samenvatting

Vanaf het moment dat Darwin zijn boek *De uitdrukking van Emoties bij Mens en Dier [The Expression of the Emotions in Man en Animals]* uitgaf in 1872, heeft een niet aflatende stroom publicaties over de productie en perceptie van emoties het licht gezien. Eerder onderzoek is vooral uitgevoerd onder de vlag van psychologie, fysiologie, biologie; maar verlegde allengs zijn grenzen en omvatte daarna mede de sociologie, taalwetenschap, pathologie, informatica, neurowetenschap, musicologie en tweedetaalverwerving. Daarenboven is er altijd al onderzoek gedaan naar de waarneming of uitdrukking van emoties vanuit over verschillende talen (cross-linguïstisch) en culturen (cross-cultureel) heen. Volgens sommige onderzoekers is de waarneming van emoties universeel. Andere onderzoekers menen dat de waarneming deels universeel verloopt en deels taal-of-cultuurspecifiek is. Eerder onderzoek laat zien dat emoties over het algemeen beter herkend worden als deze worden uitgedrukt door een spreker van dezelfde taal en afkomstig uit dezelfde cultuur als de luisteraar; dit wordt het in-groepvoordeel genoemd. Deze onderzoeken hebben ook uitgewezen dat automatische herkenning van door mensen geproduceerde emoties akoestische eigenschappen kan aanwijzen die menselijke luisteraars gebruiken om emoties te herkennen en als spreker mee uit te drukken. Ondanks de veelheid aan onderzoek naar de productie en perceptie van emoties, krijgen we uit het eerdere onderzoek geen goed beeld van hoe goed luisteraars van een niet-toontaal emoties kunnen waarnemen die worden uitgedrukt in een toontaal, vooral als die uitdrukking plaats heeft via het auditieve kanaal. Evenmin geeft het eerder onderzoek aan hoe succesvol niet-moedertaalsprekers van een taal vocale emoties kunnen produceren in een tweede of vreemde taal (T2), in verhouding tot wat moedertaalsprekers kunnen, vooral niet wanneer de vreemde taal (T2) een toontaal is maar de moedertaal (T1) van de spreker niet. We weten ook niet of een spreker vocale emoties in zijn T2 even goed kan uitdrukken als in zijn moedertaal (T1).

Daarom was het eerste doel van dit proefschrift om experimenteel te onderzoeken hoe goed Chinese moedertaalsprekers en Nederlandse niet-moedertaalsprekers van het Mandarijn (Standaard Chinees) vocaal uitgedrukte emoties kunnen waarnemen in een toontaal, in casu het Mandarijn. De niet-moedertaalluisteraars waren hetzij naïeve luisteraars hetzij gevorderde Nederlandse leerders van het Mandarijn (met dezelfde moedertaalachtergrond als de naïeve luisteraars). In de tweede plaats heb ik onderzocht of Nederlandse leerders van het Mandarijn even goed in staat zijn vocale emoties te produceren in de vreemde taal als in hun eigen niet-toontaal, d.w.z. het Nederlands; tegelijkertijd heb ik nagegaan hoe goed naïeve moedertaalluisteraars en gevorderde leerders van het Mandarijn vocale emoties kunnen waarnemen die worden uitgedrukt door L2-sprekers van de toontaal. Naderhand heb ik een akoestische analyse uitgevoerd om de vocale correlaten op te sporen die sprekers en luisteraars gebruiken bij resp. de productie en perceptie van de vocale emoties. Als laatste heb ik onderzocht of het in-groepvoordeel dat door andere onderzoekers gerapporteerd is (waarbij de claim was

luisteraars in het algemeen emoties succesvoller herkennen in hun eigen taal dan in een onbekende taal), universeel (d.w.z. taal- en cultuuronafhankelijk) is.

Vanuit theoretisch gezichtspunt probeert dit proefschrift een functionele hypothese te toetsen die claimt dat de prosodische ruimte die talen kunnen gebruiken eindig is. Als een taal bv. duur gebruikt om een segmenteel contrast te markeren tussen korte en lange klinkers, dan zal de duurparameter geen rol kunnen spelen (of hooguit een geringe rol) bij de markering van klemtoon – welk verschijnsel in andere talen sterk leunt op duuraanwijzingen (Berinstein 1979, Potisuk et al. 1999, Remijsen 2002a, b). Volgens dezelfde redenering zal een taal zoals het Mandarijn, die toonhoogteverschillen gebruikt om verschillen tussen woordvormen te signaleren (zgn. lexicale toon), minder ruimte over hebben om met behulp van toonhoogte paralinguïstischce contrasten te markeren, bv. om emoties uit te drukken. Dientengevolge voorspel ik dat moedertaal-luisteraars van het Mandarijn slechts beperkt zijn blootgesteld aan duidelijke voorbeelden van prosodisch uitgedrukte emoties. Meer algemeen voorspel ik dat luisteraars met een toontaalachtergrond minder gespitst zullen zijn op (en minder ervaren zullen zijn in) het decoderen van paralinguïstisch gebruik van prosodie dan luisteraars met een niet-toontaal als moedertaal.

In het inleidend **Hoofdstuk 1** vat ik samen wat bekend is uit eerder onderzoek en identificeer ik enkele onopgeloste vragen. Ook verschaf ik in dit hoofdstuk achtergrondinformatie over het verschil tussen toontalen en niet-toontalen, specifiek over het Mandarijn Chinees en het Nederlands, over toon en (emotionele) prosodie en over de akoestische eigenschappen van prosodie. Ik geef vervolgens een puntsgewijs overzicht van mijn onderzoeksvragen en leid specifieke voorspellingen af op grond van de functionele zienswijze die ik hierboven heb genoemd. Ook geef ik mijn beweegredenen voor de keuze van zes emotionele prosodieën die ik in deze studie gebruik ('neutraal', 'blij', 'boos', 'verbaasd', 'bedroefd' en 'sarcastisch') en stel ik haalbare onderzoeksmethoden voor om antwoorden te vonden op mijn onderzoeksvragen. Aan het eind van dit hoofdstuk schets ik de organisatie van de rest van het proefschrift.

**Hoofdstuk 2** geeft een meer gedetailleerde bespreking van eerder onderzoek. Er is veel onderzoek gedaan naar de waarneming van emotionele prosodie binnen een en dezelfde cultuurgroep. Onderzoek naar de herkenning van vocale emotie over culturen heen (zgn. cross-culturele herkenning) is slechts mondjesmaat verricht en er is nog minder vergelijkend onderzoek gedaan naar de productie van emotie. Er lijkt in het geheel geen eerder onderzoek te bestaan date en directe vergelijking maakt van de productie van emotionele prosodie in de sprekers T1 (moedertaal) en T2 (vreemde taal). In dit laatste opzicht betreedt het productiedeel van dit proefschrift onontgonnen terrein.

Veel onderzoekers hebben beweerd dat luisteraars emotionele prosodie boven kans kunnen herkennen zowel binnen hun eigen cultuur als over culturen heen. Emoties worden over het algemeen beter herkend als sprekers en luisteraars tot dezelfde culturele groep behoren. Herkenning wordt minder trefzeker naar mate de culturele afstand tussen spreker- en luisteraargroep groter is. Eerder onderzoek heeft bovendien laten zien dat succesvolle communicatie van vocale emotie zowel van de spreker als van de luisteraar afhangt, al lijkt de rol van de luisteraar van groter belang dan die van de

spreker. Sommige onderzoekers geven aan dat de productie van emotionele prosodie tot op zekere hoogte universeel (algemeen menselijk) maar dat de productie van sommige emoties eerder cultuurspecifiek is.

In dit hoofdstuk, geef ik ook information over de methodologie die ik gebruik. Het proefschrift omvat drie beoordelingsstudies. De eerste beoordelingsstudie is een luisterexperiment (Exp. 1) waarin moedertaalluisteraars van het Chinees, naïeve Nederlandse luisteraars en gevorderde Nederlandse studenten Chinees aan de hand van de prosodie (d.w.z. melodie en temporele organisatie) de boven genoemde zes Chinese emoties voortgebracht door moedertaalsprekers van het Chinees moesten beluisteren en herkennen. Dit experiment beoogt antwoord te geven op onderzoeksvraag (i) *Hoe goed kunnen Chinese moedertaalluisteraars, naïeve Nederlandse luisteraars en gevorderde Nederlandse leerders van het Chinees Chinese emotionele prosodische vormen herkennen wanneer die vocaal neergezet zijn door moedertaalsprekers van het Chinees. En wat zijn de verwarringspatronen bij elk van deze drie luisteraargroepen?* Het tweede beoordelingsexperiment viel uiteen in twee perceptie-proeven: in de eerste luisterden dezelfde drie luisteraargroepen naar dezelfde zes Chinese emotionele prosodische vormen maar die waren nu geproduceerd door gevorderde Nederlandse studenten Chinees (Exp. 2A). Dit experiment was opgezet om antwoord te vinden op onderzoeksvraag (ii) *Hoe goed kunnen Chinese moedertaalluisteraars, Nederlandse naïeve luisteraars en gevorderde Nederlandse leerders van het Chinees Chinese emotionele prosodische vormen herkennen die vocaal zijn neergezet door Nederlandse studenten Chinees. En wat zijn de verwarringspatronen bij de drie luisteraargroepen?* In de tweede perceptieproef (Exp. 2B) luisterden Nederlandse moedertaalluisteraars naar dezelfde zes emotionele prosodische vormen die nu waren geproduceerd in hun eigen taal (Nederlands) door dezelfde Nederlandse studenten Chinees. Met deze proef kon ik nagaan hoe doeltreffend dezelfde Nederlandse studenten Chinees de emotionele prosodische vormen in hun eigen taal kunnen produceren. De resultaten van Exp. 2A en Exp. 2B zijn met elkaar vergeleken om antwoord te vinden op onderzoeksvragen (iii) *Kunnen Nederlandse T2-sprekers van het Chinees emotionele prosodische vormen in the L2 even goed produceren als in hun T1 – Nederlands? Wat zijn de overeenkomsten en verschillen tussen deze twee productietypen?* en (iv) *Vormt de T2 een beperking bij de expressie van emotionele prosodie, vooral wanneer de T2-sprekers van de toontaal een niet-toontaal als moedertaal hebben. non-tonal language?* Onderzoeksvraag (v) *Is er steun voor de functionele hypothese die voorspelt dat luisteraars met een toontaalachtergrond minder gespitst zijn op (en minder ervaren zijn in) het decoderen van paralinguïstische onderscheidingen dan luisteraars met een niet-toontaal als achtergrond?* kan worden beantwoord aan de hand van het eerste en het tweede beoordelingsexperiment. Vervolgens is een akoestische analyse gemaakt van de stimuli die in de twee beoordelingsexperimenten zijn aangeboden. De resultaten van deze analyse beantwoordden onderzoeksvraag (vi) *Welke akoestische parameters dragen bij aan het onderscheid tussen de zes emotional prosodische vormen in het algemeen? Welke akoestische correlaten gebruiken sprekers en luisteraars als zij deze vocale emoties produceren en waarnemen in hun T1 en in hun T2? Passen de Nederlandse T2-sprekers van het Chinees transfer uit de moedertaal toe om emotionele prosodische vormen te produceren in het Chinees? In hoeverre weerspiegelt automatische herkenning de waarneming van de emotionele prosodische vormen door de menselijke luisteraars?* Het derde beoordelingsexperiment was symmetrisch opgezet. Het omvatte twee perceptieproeven waarin onervaren Chinese en Nederlandse luisteraars de zes emoties te horen kregen zoals die vocaal waren uitgedrukt in hun moedertaal en in de andere taal (Exp. 3). Dit experiment was opgezet om antwoord te krijgen op onderzoeksvraag (vii) *Is het in-groepvoordeel universeel, d.w.z. herkennen luisteraars emotionele*

*prosodie altijd beter als die wordt geproduceerd in hun moedertaal dan in een tweede of geheel onbekende taal? Is de herkenning van vocale emoties cross-cultureel symmetrisch bij Chinese en Nederlandse luisteraars, m.a.w. kunnen Nederlandse en Chinese luisteraars emotionele prosodische vormen in de andere taal even trefzeker herkennen?* De drie beoordelingsexperimenten tezamen geven antwoord op onderzoeksvraag (viii) *Zijn productie en perceptie van emotionele prosodie universeel? Of zijn ze eerder taal- en/of cultuurspecifiek?*

**Hoofdstuk 3** doet verslag van het eerste beoordelingsexperiment (Exp. 1), dat tevens dient als vergelijkingsconditie voor de latere proeven. Twintig Chinese moedertaalluisteraars, 20 naïeve Nederlandse luisteraars en 20 gevorderde Nederlandse L2 leerders van het Chinees moesten de zes Chinese emotionele prosodische vormen (neutraal, blij, boos, verbaasd, bedroefd en sarcastisch) herkennen. De resultaten laten zien dat de gevorderde Nederlandse T2 leerders van het Chinees de Chinese emotionele prosodie significant (54 % correct) beter herkenden dan de Chinese moedertaalluisteraars zelf (46 % correct) en ook beter dan Nederlandse naïeve luisteraars (46 % correct). De resultaten laten ook zien dat naïeve T2 (Nederlandse) luisteraars emoties in een onbekende taal (Mandarijn) even goed kunnen herkennen als de moedertaalluisteraar zelf. De Chinese moedertaalluisteraars gaven dus geen blijk van een in-groepvoordeel: zij identificeerden de Chinese emoties niet trefzekerder en met meer zelfvertrouwen. Voor alle drie de luisteraargroepen was 'neutraal' de gemakkelijkste emotie om te identificeren; 'boos' werd door alle luisteraargroepen even vaak correct herkend. De voorspelling die ik gedaan heb aan het begin van mijn onderzoek is hiermee bevestigd: luisteraars met een toontaalachtergrond zijn minder bedreven in het paralinguïstisch gebruik van prosodie dan luisteraars die geen toontaal als achtergrond hebben. De resultaten van dit hoofdstuk vormen de nullijn waartegen de resultaten van het volgende hoofdstuk tegen zullen worden afgezet.

**Hoofdstuk 4** onderzoekt de verschillen in perceptie van zes Chinese emotionele prosodische vormen (neutraal, blij, boos, verbaasd, bedroefd en sarcastisch) die werden geproduceerd door Nederlandse T2-sprekers van het Chinees en die welke waren ingesproken door moedertaalsprekers van het Chinees (controlegroep). Dit hoofdstuk vergelijkt de resultaten van Exp. 1 en Exp. 2A. Twintig Chinese T1-luisteraars, 20 naïeve Nederlandse T2-luisteraars en 20 gevorderde Nederlandse studenten Chinees beluisterden de Chinese emotionele prosodische vormen die waren uitgedrukt door zowel T1- als T2-sprekers van het Chinees. De resultaten laten zien dat de drie luisteraargroepen de emotionele prosodische vormen die waren ingesproken door Chinese moedertaalsprekers significant beter (49% correct) herkenden dan die welke waren geproduceerd door L2-sprekers van het Chinees (39% correct). Ook blijkt dat de naïeve T2-luisteraars de emoties in de onbekende taal even goed konden herkennen (46% correct) als de Chinese moedertaalluisteraars zelf. Hoewel de gevorderde Nederlandse studenten van het Chinees het iets beter deden dan de twee overige luisteraargroepen, was er geen significant verschil tussen de drie luisteraargroepen in de identificatie van emotionele prosodie die is geproduceerd door T2-sprekers. De functionele voorspelling is hiermee opnieuw bevestigd: luisteraars met ene toontaalachtergrond zijn minder bedreven in het paralinguïstisch gebruik van prosodie dan luisteraars met een niet-toontaalachtergrond. of a non-tonal language. Er zijn dus situaties waarin het taalkundig gebruik van een akoestische eigenschap beperkingen oplegt aan het gebruik van die eigenschap ten behoeve van de communicatie van emotie.

**Hoofdstuk 5** presenteert een completer beeld van hoe Nederlandse T2 sprekers van het Chinees de zes emotionele prosodische vormen in de T2 (Chinees) produceren en hoe zij dezelfde emotionele prosodische vormen in hun moedertaal (Nederlands) uitdrukken. Dit hoofdstuk rapporteert de resultaten van het eerste en het tweede beoordelingsexperiment op een geïntegreerde wijze vanuit het oogpunt van de productie. De resultaten laten zien dat emotionele prosodische vormen geproduceerd door Nederlandse T2-sprekers van het Chinees over de gehele linie minder herkenbaar zijn (39% correct door Chinese luisteraars) in hun T2 dan die welke door Chinese moedertaalsprekers werden uitgesproken (46% correct). Nederlandse T2-sprekers van het Chinese slaagden er beter in vocale emoties te produceren in hun moedertaal (57% correct geïdentificeerd door Nederlandse luisteraars). De voorspelling die ik aan het begin van dit hoofdstuk heb geformuleerd, is hiermee bevestigd: spreken in een vreemde taal beperkt de sprekers expressie bam emotie. De resultaten laten ook zien dat de naïeve Nederlandse luisteraars in staat waren de emoties in de onbekende taal even goed te herkennen als de moedertaalluisteraars van het Mandarijn zelf. Bovendien lieten de naïeve Nederlandse luisteraars een in-groepvoordeel zien: zij identificeerden de emoties in hun eigen taal (Nederlands) accurater dan die in het Chinees.

**Hoofdstuk 6** bevat een akoestische analyse van drie typen productie van emotionele prosodie: T1-Mandarijn, T2-Mandarijn en T1-Nederlands (de twee laatste typen waren geproduceerd door dezelfde individuen). Acht akoestische correlaten zijn onderzocht: spreeksnelheid (tempo), gemiddelde grondfrequentie (toonhoogte of F0), de standaard deviatie van de toonhoogte (SD_F0), snelheid van F0-verandering (helling_F0), compactheid van de spectrale energieverdeling, standaard deviatie van de intensiteit (SD_int), jitter (jengel, d.w.z. de cyclus-op-cyclusvariatie van de periode van de stembandpulsen, als maat voor stemvastheid) en HNR (harmonischen-ruisverhouding, als maat voor de ruizigheid van de stembandpulsen). Ik heb ook een automatische herkenning van de zes emotionele prosodische vormen zoals geproduceerd door de drie sprekergroepen uitgevoerd met behulp van Lineaire Discriminant Analyse (LDA). De akoestische analyse toont aan dat de grondfrequentieparameters, d.w.z. gemiddelde F0, SD_F0 en F0-helling, van grote invloed zijn bij de productie van vocale emotie door de drie sprekergroepen. Deze bevinding komt overeen met een onderzoek van Scherer (1996), waarin eveneens is gevonden dat F0 een cruciale rol speelt bij de productie van emotionele prosodie. Jitter en standaard deviatie van de intensiteit dragen weinig bij aan het onderscheid tussen de emoties in mijn onderzoek. 'Tempo' en 'compactheid' zijn alleen onderscheidend bij moedertaalsprekers van het Mandarijn. Helling van de F0 wijst uit dat Chinese sprekers een stijgende intonatie gebruiken om verbazing uit te drukken, wat eerdere studies bevestigt die stelden dat veel toontalen verbazing uitdrukken met stijgende toonhoogte (Yip 2006). Daarenboven onderscheidt HNR duidelijk 'bedroefd' van 'neutraal' in T2-Mandarijn en in T1-Nederlands (geproduceerd door dezelfde Nederlandse individuen) maar niet in T1-Mandarijn. Samenvattend: grondfrequentie is een heel invloedrijke variabele in de productie en perceptie van vocale emotie in het algemeen. Andere parameters die in dit hoofdstuk onderzocht zijn, dragen eveneens bij aan het onderscheid tussen emotionele prosodische vormen, maar deze zijn meer emotiespecifiek en/of sprekerspecifiek.

De resultaten van de LDA geven aan dat de emotionele prosodie die door menselijke sprekers is voortgebracht, automatisch ruim boven kans herkend kan worden aan de

hand van de uitgevoerde akoestische metingen (totaal 50% correcte identificatie). Er blijkt een significante correlatie tussen de perceptieve verwarringen van emotionele categorieën bij de menselijke luisteraars en die van de automatische herkenning in het huidige onderzoek. Dit geeft aan dat de automatische herkenning tot op zekere hoogte een model is van de menselijke perceptie van deze emoties. Dit sluit echter geenszins uit dat ook nog andere akoestische parameters een bijdrage kunnen leveren aan de productie en de perceptie van emotionele prosodie in het algemeen, die wij in deze dissertatie niet heb kunnen identificeren.

De akoestische analyse wijst uit dat de Nederlandse T2-sprekers bij de productie van emotionele prosodie in hun T2 (Chinees) sommige akoestische parameters op dezelfde manier gebruiken als in hun T1 (Nederlands), bv. SD_F0 en SD_Int. We kunnen daarom de conclusie trekken dat T2-sprekers T1-transfer gebruiken als strategie om vocale emoties te produceren in de T2. Deze strategie zal echter niet voor alle emoties het gewenste resultaat opleveren, bv. niet voor 'verbaasd' en 'sarcastisch'. De resultaten suggereren bovendien dat de T2-sprekers niet automatisch de signalering van emotionele prosodie overnemen van voorbeelden in de doeltaal. Het lijkt er eerder op dat deze gevorderde T2-sprekers van het Chinees een hydride systeem hebben ontwikkeld om emotionele prosodie in de T2 te produceren. Dat hybride systeem benadert de manier waarop moedertaalsprekers van het Chinees vocale emoties uitdrukken, bv. de manier waarop tempo, gemiddelde F0, helling van de F0, compactheid en jitter worden ingezet, maar houdt wat betreft variabiliteit in F0 en intensiteit vast aan wat gebruikelijk is in de moedertaal van de T2-sprekers.

**Hoofdstuk 7** onderzoekt de waarneming over en weer van emotionele prosodie die wordt moedertaalluisteraars en door ongeoefende T2-luisteraars (Exp 3). Twintig Chinese en 20 Nederlandse moedertaalluisteraars zonder enige eerdere ervaring met elkaars taal, hadden de opdracht om de zes emotionele prosodische vormen in deze twee talen te identificeren. De resultaten geven aan dat onervaren Nederlandse luisteraars (46% correct) de emotionele prosodie in de onbekende taal (Chinees) even goed konden herkennen als de Chinese moedertaalluisteraars zelf (46% correct); en zij voerden hun taak significant beter als ze dezelfde emotionele prosodieën moesten herkennen in hun moedertaal (Nederlands, 57% correct). De Chinese onervaren luisteraars daarentegen waren alleen in staat om redelijk goed de emotionele prosodie te herkennen als die gesproken was in hun eigen taal (46% correct) maar slaagden er niet in om de vocale emoties boven kans te herkennen als die werd uitgedrukt in de voor hun onbekende taal (Nederlands, 15% correct). Dit resultaat bevestigt het bestaan van het in-groepvoordeel dat ook door andere onderzoekers gerapporteerd is en dat erop neerkomt dat luisteraars emotionele prosodie in het algemeen beter herkennen als die is geproduceerd door sprekers van hun eigen taal dan door sprekers van een onbekende taal. De resultaten suggereren bovendien dat de perceptie van vocale emotie cross-cultureel niet symmetrisch is, zodat de ene groep luisteraars (afhankelijk van taal en/of cultuur) over de hele linie beter is in het identificeren van emotionele prosodie dan de andere groep. Dit ondersteunt op zijn beurt de functionele hypothese die voorspelt dat luisteraars met een toontaalachtergrond in het algemeen minder bedreven zullen zijn in de perceptie van emotionele prosodie dan luisteraars die een niet-toontaal als moedertaal hebben.

**Hoofdstuk 8** zet de belangrijkste bevindingen van dit proefschrift nog eens op een rij en probeert dan aan de hand daarvan antwoord te geven op de onderzoeksvragen die in hoofdstuk 1 gesteld zijn. Het hoofdstuk wordt afgesloten met een bespreking van aspecten die in toekomstig onderzoek verbeterd zouden kunnen worden.

# 摘要

自达尔文 1872 年出版了他的《人类和动物的情绪表达》之后，学术界出现了非常多的关于情感认知和生成方面的研究。早期的研究主要集中在心理学、生理学及生物学领域，但后期的研究则扩展到其它诸多领域，比如：社会学、语言学、病理学、计算机科学、神经学、音乐学及第二语言习得。除此以外，很多研究人员一直在从事关于跨文化跨语言方面的情感认知和生成的研究。一些学者认为，情感的认知是具有普遍性的。但是，一些学者则认为，一部分情感的认知是具有普遍性的，一部分却是基于说话人-听话人所具有的特定文化语言背景的。前人的研究表明，如果听话人和说话人是来自相同语言文化圈的，那么他们之间的情感识别率会比那些来自非同一语言文化圈的听话人和说话人的识别率高（这种现象也称作，“同一群体优势”)。以前的研究还表明，计算机自动识别系统可以显示一些人类认知和生成情感的声学线索。虽然前人在情感认知和生成方面的研究硕果磊磊，但是前人的研究并没有解决所有的问题，比如 “母语为非声调语言的听话人如何识别用声调语言表达的情感?”；“若说话人的母语为非声调语言，但其外语为声调语言，和声调语言母语者相比，此说话人应该如何用他的外语（即声调语言）表达情感?”；“外语学习者可否像用其母语表达情感一样，用其二语顺畅地表达情感?” 这些问题都是有待解决的研究问题。

因此，本论文的第一个主旨就是用实验的方法研究母语为中文的听话人和母语为非中文的听话人 (比如，荷兰人) 可以在何种程度上正确识别用中文表达的情感。本论文中，母语为非中文的听话人包括完全不懂中文的荷兰人和高级汉语学习者 (荷兰人)。第二个主旨是，本研究要调查汉语作为外语的荷兰学习者是否可以用其第二语言 (中文) 表达情感，如同他们用自己的母语（荷兰语) 表达情感那样。同时，本博士课题还要研究，完全不懂中文的荷兰人和汉语作为外语的荷兰学习者 (高级水平) 可以在何种程度上正确识别由荷兰汉语学习者用中文表达的声音情感。之后笔者将进行声学分析，这会让我们了解听话人和说话人在情感认知和生成过程中使用了哪些声学线索。最后，本论文要调查前人提出的“同一群体优势” 是否具有普遍性。(前人阐述：在情感是用听话人的母语和其完全不懂的外语表达的情况下，听话人更会准确识别用自己母语所表达的情感。)

从理论角度看，本论文的另一个目标是检验一个 “功能假说”。此假说认为，语言的音律空间是有限的。比如，和一些完全使用音长来表达重音的语言相比 (贝瑞斯坦 1979 年，泊蒂徐克 1999 年，乐麦森　2002 年 a,b)，如果一种语言已

经使用音长来区别长元音和短元音，那么音长这个参数就不会被用来 (或不会被完全用来) 表达重音。同样的道理，汉语用声调来表达词汇意义，声调会占用汉语的音律空间，所以可以用来表达超音短成分 (比如，情感韵律) 的剩余空间则会减少。这也就是说，声调语言会对情感韵律的表达有压制作用。因此，笔者大胆假设，和非声调语言的母语者 (比如，荷兰人或英国人) 相比, 中文母语者对于只用声音表达的情感的感知是不清晰的。更宽泛地假设，和母语为非声调语言的听话人相比，母语为声调语言的听话人不会对只通过超音短成分表达的情感有更为准确的识别。实际上，他们缺乏只用声音表达情感的经验。

本论文的第一章为概论。本章总结了前人的研究成果并给出了前人未解决的问题。本章还提供了关于声调语言和非声调语言的背景知识 (具体是，汉语和荷兰语)、声调和情感韵律、以及情感韵律声学方面的相关知识。除此以外，本章分条列出了此博士论文的研究问题和由前文提及的 “功能假说” 所引发的具体假设。并且，本章陈述了选择当前六种目标情感 (“中性”、“高兴”、“生气”、“吃惊”、“伤心” 及 “讽刺”) 进行研究的原因。同时，本章亦提供了相应的可行性研究方法。此章的最后列出了论文的通篇结构。

本论文的第二章详细地回顾了前人的研究。前人大多在同一语言文化群体范畴内展开情感认知的研究，但在跨文化范畴内展开的情感认知研究则相对较少, 而在跨文化范畴内展开的情感生成方面的研究就更为罕见了。至本论文截稿, 笔者尚未找到直接探索关于 “说话人如何用母语及其外语表达声音情感” 的学术文章。因此，本论文中关于声音情感生成部分的研究是具有初探性的。

许多研究人员阐述：普遍来讲，不论是在同一文化圈内还是跨文化圈，听话人对于声音情感（即：情感韵律）的识别率在机会水平之上。但是，研究表明, 如果说话人和听话人来自同一文化群体，那么他们之间的情感识别率比较高。若说话人和听话人的背景文化差别越大，他们之间的情感识别率就越低。研究还显示，虽然情感交流更依赖于听话人，但是成功的声音情感交流还是依靠说话人和听话人的共同努力。一些学者认为，情感韵律的跨文化生成可能在某种程度上是具有普遍性的；但是一些情感的生成则是具有特定文化属性的。也就是说，不同语言文化背景的说话人表达情感的方式不同。

在第二章中，笔者也提供了研究方法，详细内容如下：本论文一共包括三个识别研究。第一个识别研究只包括一个认知实验 (实验 1)：在实验 (1)中，中文母语听话人、完全不懂中文的荷兰听话人和荷兰高级汉语学习者识别了前文提到的六种情感。这六种情感是由母语为中文的说话人表达的。这个实验的目的是回答第一个研究问题 (i) “在何种程度上，中文母语听话人、完全不懂中文的荷兰听话人和荷兰高级汉语学习者可以正确识别由中文母语说话人表达的中文

情感韵律？这三组被试的情感混乱格局是什么?”第二个识别研究包括两个识别实验 (2A 和 2B)：在第一个识别实验 (2A) 中，前面所述的三组被试识别了由汉语作为外语的荷兰说话人所表达的汉语情感韵律。这个实验旨在回答研究问题 (ii) “在何种程度上，中文母语听话人、完全不懂中文的荷兰听话人和荷兰高级汉语学习者可以正确识别由汉语作为外语的荷兰说话人表达的中文情感韵律？这三组被试的情感混乱格局是什么?”在第二个识别实验 (2B) 中，母语为荷兰语的听话人识别了前面所述的六种情感。所不同的是，在此次实验中，前文提及的六种情感是用荷兰语表达的，且是由实验 (2A) 中的那些汉语作为外语的荷兰说话人所表达的。这个实验是为了研究汉语作为外语的荷兰说话人可以在何种程度上用自己的母语 (荷兰语) 表达情感韵律。实验 (2B) 的结果会和实验 (2A) 的结果进行比较，从而回答研究问题 (iii) “汉语作为外语的荷兰说话人可否用其第二语言 (中文) 表达情感韵律，这种表达是否和用他们的母语 (荷兰语) 表达得一样好？这两种表达有什么相同之处和不同之处？”此外，比较结果还会回答研究问题 (iv) “是否外语会限制情感韵律的表达，特别是会限制那些母语是非声调语言但其外语是声调语言的外语言学习者?”在完成第一个识别研究和第二个识别研究之后，笔者将回答研究问题 (v) “ ‘功能假说’ 所提出的假设——‘和母语为非声调语言的听话人相比，母语为声调语言的听话人不会对只通过超音短成分表达的情感有更为准确的识别。实际上，他们缺乏只用声音表达情感的经验。’ ” 进而我们可以清楚地知道前面提及的 “功能假说” 是否正确。笔者将会对在第一个识别研究和第二个识别研究中所用到的语料进行声学分析。声学分析结果会回答研究问题（vi）哪些声学线索可以用来区分不同的情感韵律？说话人和听话人在生成和识别情感韵律的过程中使用了哪些声学线索？汉语作为外语的荷兰说话人是否在用其外语（汉语）表达情感韵律时使用了“母语转移”的话语策略？在何种程度上计算机自动识别系统可以客观反映人类说话人所使用的声学线索？第三个识别研究采取了交互识别的方法。这个识别研究包括两个识别实验 (3A 和 3B，通称“实验 3”)。在这两个识别实验中，完全不懂荷兰语的中文母语听话人和完全不懂中文的荷兰母语听话人交互识别了由中文和荷兰文母语者用其母语 (中文和荷兰文) 所表达的前面所提的六种情感韵律。这两个交互识别实验是为了验证由前人提出的 “同一群体优势” 的普遍性，以此回答研究问题 (vii) “ ‘同一群体优势’ 是否具有普遍性？如果是，这也就是说，与识别用听话人的外语或其完全不懂的语言表达的情感相比，听话人更准确地识别用自己母语表达的声音情感。声音情感的认知在跨文化的范畴内是否对称？如果是，这也就是说，中文母语听话人和荷兰母语听话人在跨文化范畴内应该具有同等的情感识别能力。” 第一、第二和第三个识别研究会共同回答研究问题 (viii) “情感韵律的认知和生成是否具有普遍性？他们是否更是由说话人－说话人具有的特定语言和文化背景决定的?”

第三章报告了第一个识别研究 (实验 1) 的结果。这组结果将被用作之后研究的基准线。二十名中文母语听话人，20 名完全不懂中文的荷兰母语听话人和 20 名荷兰高级汉语学习者识别了前文所提的六种中文情感韵律 (“中性”、“高兴”、“生气”、“吃惊”、“伤心” 及 “讽刺”)。结果显示，荷兰高级汉语学习者

对这些中文情感韵律的识别率 (54% 正确) 显著高于中文母语听话人 (46% 正确) 和完全不懂中文的荷兰母语听话人 (46% 正确) 的识别率。实验结果还显示，完全不懂中文的荷兰母语听话人可以和中文母语听话人一样正确识别用中文表达的情感。此次实验中，中文母语听话人没有体现出 “同一群体优势”, 这也就是说，中文母语听话人并没有更准确更自信地识别用其母语 (中文) 表达的声音情感。“中性” 情感韵律对于这三组被试来说，是最容易识别的情感; 并且，这三组被试都比较准确地识别了 “生气” 情感韵律。前面所提出的假设在此被证实：母语为声调语言的听话人不会对只通过超音短成分表达的情感有更为准确的识别。此次实验的结果为第四章提供研究基准线。

第四章对比了前面提及的三组听话人对汉语作为外语的荷兰说话人和对中文母语说话人所表达的六种中文情感韵律 (“中性”、“高兴”、“生气”、“吃惊”、“伤心” 及 “讽刺”) 的认知。这一章报告了实验 (1) 和实验 (2A) 的实验结果。二十名中文母语听话人，20 名完全不懂中文的荷兰母语听话人和 20 名荷兰高级汉语学习者分别识别了由母语说话人和二语说话人所表达的六种中文情感韵律。实验结果显示，三组听话人对由母语说话人表达的中文情感韵律的识别率 (49% 正确) 显著高于对由外语说话人表达的中文情感韵律的识别率 (39% 正确)。并且，完全不懂中文的荷兰听话人 (46% 正确）可以和中文母语听话人一样正确识别用中文表达的情感。在对外语说话人表达的中文情感韵律的识别过程中，虽然相对于中文母语听话人和完全不懂中文的荷兰听话人，荷兰高级汉语学习者表现出稍高的正确识别率，但这三组被试中没有一组被试的识别率显著高于其它两组被试的。“功能假说” 在此次实验中再次被证实：和母语为非声调语言的听话人相比，母语为声调语言的听话人不会对情感韵律 (超音短成分) 有较高的识别率。这也就是说，在一些情况下，一种语言如果在音律空间里使用特定一种声学特征来实现话语表达 (比如，汉语使用声调来表达词义), 那么它将压制该语言在声音情感方面的表达。换句话说，声调的使用很可能压制情感韵律的表达。

第五章是从情感生成的角度撰写的，完整地报告了第一个识别研究和第二个识别研究的结果：汉语作为外语的荷兰说话人如何用其二语 (中文) 表达前面所提的六种情感韵律；并且他们是如何用自己的母语 (荷兰语) 表达这些情感韵律的。结果显示，整体来说，汉语作为外语的荷兰说话人表达的中文情感韵律的识别率 (39% 正确，中文母语听话人) 比那些汉语母语者表达的中文情感韵律 (46% 正确，中文母语听话人) 的识别率要低。结果显示，汉语作为外语的荷兰说话人更准确地用自己的母语荷兰语来表达声音情感 (57% 识别率，荷兰母语听话人)。在本章开始时所提出的假设被证实：外语限制外语说话人的声音情感表达。本章结果还显示，完全不懂中文的荷兰听话人可以和中文母语听话人一样正确识别用中文表达的情感。并且，完全不懂中文的荷兰母语听话人在识别用自己母语荷兰语表达的六种情感韵律时表现出 “同一群体优势”。这

也就是说，完全不懂中文的荷兰母语听话人更准确地识别了用自己母语荷兰语表达的情感韵律。

第六章报告了三类情感韵律生成的声学分析。这三类情感韵律为：由汉语母语说话人表达的中文情感韵律，由汉语作为外语的荷兰说话人表达的中文情感韵律以及由汉语作为外语的荷兰说话人用其母语荷兰语表达的荷兰情感韵律。本章分析了八个声学线索："音长"、"平均基频" (又称，音高或 F0)，"基频标准差" (SD_F0)，"基频变化率" (slope_F0)，"声音频谱分布的致密性"，"音强的标准差"，"抖动" (jitter，声门脉冲的周期到周期变化) 以及"谐波和噪声成份的比值" (HNR)。在此章中，笔者采用了计算机自动识别程序对这三类情感韵律进行了自动识别，具体使用的是 "线性判别分析" (LDA)。声学分析显示，在这三类情感韵律生成中，"基频"，包括 "平均基频"、"基频标准差" 和 "基频变化率" 是非常有影响力的声学线索。这个发现证实了谢埃尔 (1996 年) 的研究。谢埃尔认为，"基频" 在情感韵律的生成过程中发挥重要作用。在本博士研究中，"抖动" (jitter) 和 "音强的标准差" 并没有在区别情感韵律上发挥多少作用。"音长" 和 "声音频谱分布的致密性" 这两个参数只对由汉语母语说话人表达的中文情感韵律表现灵敏。"基频变化率" (slope_F0) 显示，中文母语说话人用上扬句调来表达吃惊的情感韵律。这个发现证实了前人的研究：很多声调语言用上扬句调表达吃惊（叶普，2006 年)。并且，在汉语作为外语的荷兰说话人用其外语中文表达中文情感韵律和他们用其母语荷兰语表达荷兰情感韵律时，"谐波和噪声成份的比值" (HNR) 这个参数非常清晰地把 "伤心" 从 "中性" 情感中分辨出来。但在汉语母语说话人表达中文情感韵律过程中，这个参数显示不敏感。综上所述，一般来说，"基频" 在情感韵律生成过程中是一个非常有影响力的参数。本章研究的其它七个参数都在不同程度上起到了区分情感韵律的作用，但他们更多则是由具体情感或说话人类型决定的。

"线性判别分析" (LDA) 的结果显示，计算机自动识别系统可以基于上述八个声学参数识别由自然人所表达的情感韵律。整体识别率 (50% 正确) 在机会水平之上。在本研究中，计算机自动识别系统的情感混乱格局和自然听话人的混乱格局有显著相关性，这就意味着，计算机自动识别系统可以在一定程度上反映自然听话人对声音情感的认知。但是，一般来说，除了本论文研究的8个声学线索外，还会有其它一些声学线索在情感韵律的生成和认知过程中起到作用。但本论文未对这些线索有所涉及。

声学分析显示，汉语作为外语的荷兰说话人在用其二语 (中文) 表达情感韵律的过程中，运用了母语（荷兰语）的情感韵律表达法。这体现在"基频标准差"(SD_F0) 和 "音强的标准差" 两个参数上。因此，我们可以得出一个结论："母语转移" 对于外语说话人用其外语表达声音情感来说，是一个重要的话语策略。但是这个策略并不是对所有情感都适用，比如，这个策略就并不适用于 "吃惊" 和 "讽刺" 的表达。研究结果显示，外语学习者不能在其外语学习过程中自动习得母语者表达情感的方式方法。貌似高级外语学习者自己生成

了一种特殊的 “混合系统”，他们用这种系统来表达外语情感韵律。这种 “混合系统” 和中文母语者表达声音情感的方式有一定相似之处 (这在 “音长”、“平均基频”、“基频变化率”、“声音频谱分布的致密性” 以及 “抖动” 参数上有所体现)。但在另一些方面，比如 “基频变化范围” 和 “音强” 等参数上, 高级外语学习者则选择用自己母语的表达方式来用进行外语情感韵律的表达。

第七章调查了母语者和非母语者对于情感韵律的认知情况。这个认知实验 (3) 是用交互识别的方法展开的。二十名完全不懂荷兰文的中文母语听话人和 20 名完全不懂中文的荷兰母语听话人互相识别了用中文和荷兰文表达的情感韵律。结果显示，完全不懂中文的荷兰母语听话人可以和中文母语者一样正确识别用中文表达的情感，正确识别率为 46%。他们在识别用自己母语荷兰语表达的情感韵律时，识别率显著提高（正确识别率 57%)。与次相悖的是，完全不懂荷兰文的中文母语听话人只能识别用自己母语中文表达的情感韵律 (正确识别率 46%)，却不能在机会水平之上识别用荷兰文表达的声音情感 (正确识别率仅为 15%)。这个发现证实了前人所提出的 “同一群体优势” 的普遍性: 一般来说，听话人更能正确地识别用自己母语表达的声音情感，而不是用一种完全不懂语言表达的声音情感。结果还显示，在跨文化声音情感认知的范畴内, 不同文化群体对于用其它语言表达的声音情感的识别能力是不对称的。这也就是说，一些文化群体会比另一些文化群体在识别声音情感方面表现得好。这个发现再次印证了前面所提的 “功能假说”：母语为非声调语言的听话人会比母语为声调语言的听话人在识别声音情感方面表现得好。

第八章回顾了本论文的所有重要发现，并回答了第一章中所提出的所有研究问题。这一章的最后对一些尚未解决的问题进行了讨论。这些讨论会对今后的研究提供一些参考。

# Appendices

**Appendix 1 (a). Instruction and answering card for the perception experiment of the Chinese emotional prosody by native Chinese listeners (instructed in Chinese).**[20]

## 中文情感语句听力感知测试

**请您如实填写下列内容。**

**姓别：**
**年龄：**
**母语：**

---

**解释说明**：此次试验中，您将听到 144 个中文情感语句，共包括 6 种情感：中性（无情感的陈述句），高兴，生气，吃惊，伤心和讽刺。请您在听到一个情感语句后马上从所给选项中做出选择，并且对所听到的情感句子做出可信度评分。可信度分值为：3—2—1。每题只有一个正确答案，所以不可多选或不选。但可信度分值请仅凭您的感觉评断，其没有正确答案。句与句之间间隔时间为 6 秒。

可信度评分：3： 说话人非常好地表达了情感，我很肯定我的答案
2： 说话人一般地表达了情感，我不太肯定我的答案
1： 说话人完全没有适当地表达情感，我完全凭猜测来辨别

---

[20] This answering card was used for the perception experiment of the Chinese emotional prosody produced by both Chinese native and Dutch L2 speakers of Chinese. In the perception of L2-produced Chinese emotional prosody there were only 96 choices, since two sentences were discarded. It was also re-used by the Chinese novice listeners in the third judgment study with another title (in Chinese): Perception experiment of emotional prosody in an unknown language.

**例子：**

| | 中性 | 高兴 | 生气 | 吃惊 | 伤心 | 讽刺 | 可信度? |
|---|---|---|---|---|---|---|---|
| 1. | ☐ | ☐ | ☐ | ☑ | ☐ | ☐ | 3 |
| 2. | ☐ | ☑ | ☐ | ☐ | ☐ | ☐ | 2 |

## 中文情感语句听力感知测试答题卡

| | 中性 | 高兴 | 生气 | 吃惊 | 伤心 | 讽刺 | 可信度? |
|---|---|---|---|---|---|---|---|
| 1. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| 2. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| 3. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| 4. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| 5. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| 6. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| 7. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| 8. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| 9. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| 10. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| | 中性 | 高兴 | 生气 | 吃惊 | 伤心 | 讽刺 | 可信度? |
| … | | | | | | | |
| 143. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| 144. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |

**谢谢您的参与！**

**Appendix 1 (b). Instruction and answering card for the perception experiment of the Chinese emotional prosody produced by native Chinese (instructed in listeners' native language - Dutch).**[21]

## Waarnemingsexperiment:

## Vocale emotie in een vreemde taal

**Uw geslacht:** ☐ **man** ☐ **vrouw**

**Uw leeftijd:** _____

**Uw moedertaal:** _____

---

**Instructie**

U hoort aanstonds 144 korte zinnen in het Chinees die een emotie of attitude (standpunt van de spreker ten opzichte van de verbale inhoud van de boodschap) uitdrukken. Er zijn zes beoogde emoties/attitudes: *neutraal* (geen emotie of attitude), *blij*, *boos*, *verbaasd*, *verdrietig* en *sarcastisch*. Direct nadat u een zin heeft gehoord, maakt u een keuze welke emotie u denkt dat de spreker probeerde uit te drukken. U moet hierbij kiezen uit een van de zes gegeven mogelijkheden. U mag niet twee of meer emoties kiezen of een regel blanco laten. De zes mogelijkheden worden op het antwoordvel steeds duidelijk aangeven

Tevens vragen wij u in de meest rechtse kolom op het antwoordblad steeds met een cijfer aan te geven hoe zeker u bent van uw keuze:

---

[21] This answering card was used by both the naïve Dutch listeners and advanced Dutch learners of Chinese for the perception experiment of the Chinese emotional prosody produced by Dutch L2 speakers of Chinese, but only with 96 choices, since two sentences were discarded in this experiment. It was also re-used by the Dutch native listeners in the third judgment study with another title: Waarnemingsexperiment: Vocale emotie in Nederlandse taal.

Cijfer 3: De spreker heeft de emotie zeer sterk uitgedrukt; ik ben heel zeker van mijn antwoord.
Cijfer 2: De spreker heeft de emotie matig duidelijk uitgedrukt; ik ben matig zeker van mijn antwoord.
Cijfer 1: De spreker heeft de emotie zwak uitgedrukt; ik kon eigenlijk slechts gissen.

U heeft per zin zes seconde de tijd om uw keuze te maken en een cijfer in te vullen.

Voorbeelden:

| | **Neutraal** | **Blij** | **Boos** | **Verbaasd** | **Verdrietig** | **Sarcastisch** | **Zekerheid?** |
|---|---|---|---|---|---|---|---|
| 1. | ☐ | ☐ | ☐ | ☑ | ☐ | ☐ | 3 |
| 2. | ☐ | ☑ | ☐ | ☐ | ☐ | ☐ | 2 |

**Antwoordblad in het Nederlands**

| | Neutraal | Blij | Boos | Verbaasd | Verdrietig | Sarcastisch | Zekerheid? |
|---|---|---|---|---|---|---|---|
| 1. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| 2. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| 3. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| 4. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| 5. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| 6. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| 7. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| 8. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| 9. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| 10. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| | Neutraal | Blij | Boos | Verbaasd | Verdrietig | Sarcastisch | Zekerheid? |
| 11. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| 12. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| 13. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| 14. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| 15. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| … | | | | | | | |
| 143. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| 144. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |

**Einde van het experiment. Dank u wel voor uw medewerking.**

**Appendix 1 (c). Instruction and answering card for the perception experiment of the Chinese emotional prosody by native Chinese listeners (English translation).**

## Perception Experiment of Chinese Emotional Prosody[22]

**Your sex:** ☐ **male** ☐ **female**

**Your age:**

**Your mother tongue:**

---

**Instruction**

You will hear 144 short sentences expressed in different emotions in this perception experiment. There will be six emotions in the experiment: neutral (no emotion), happy, angry, surprised, sad and sarcastic. Please choose the intended emotion of the speaker from the six given emotions immediately after you hear a stimulus. There will be only one correct answer for each sentence, so you are not allowed to choose more than one emotion or leave an answer empty. Meanwhile, you are asked to give your confidence rating for your answer for each sentence. The confidence rating is a 3-2-1 scale. There is no correct answer for the confidence rating. There will be six seconds pause between the stimuli for you to choose a correct answer.

Confidence rating:

**Confident** (3): The speaker expressed the intended emotion well. I am very confident in my answer'

**Middle confident** (2): The speaker reasonably expressed the intended emotion. But I am not so sure about my answer.

**Not confident** (1): The speaker did not express the intended emotion very well. I made the choice only by guessing.

[22] The English title for Appendix 1(b) is *Perception Experiment of Emotional Prosody in a Foreign Language*. Since the answering sheet was originally designed for both native and non-native listeners of Chinese, the titles were slightly different.

**Examples:**

| | Neutral | Happy | Angry | Surprised | Sad | Sarcastic | Confidence? |
|---|---|---|---|---|---|---|---|
| 1. | ☐ | ☐ | ☐ | ☑ | ☐ | ☐ | 3 |
| 2. | ☐ | ☑ | ☐ | ☐ | ☐ | ☐ | 2 |

**Answering sheet in English**

| | Neutral | Happy | Angry | Surprised | Sad | Sarcastic | Confidence? |
|---|---|---|---|---|---|---|---|
| 1. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| 2. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| 3. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| 4. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| 5. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| 6. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| 7. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| 8. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| 9. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| 10. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| | Neutral | Happy | Angry | Surprised | Sad | Sarcastic | Confidence? |
| 11. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| … | | | | | | | |
| 143. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| 144. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |

**Thank you for your participation!**

**Appendix 2. Subgroups differentiated by the eight acoustic parameters: tempo, F0_mean, SD_ F0, slope of F0, compactness, SD_intensity, jitter and HNR (Bonferroni post-hoc procedure).**

**Appendix 2.1 Subgroups differentiated by tempo.**

**Zscore: Tempo Language = Dutch**

| | | Subset for alpha =.05 | |
|---|---|---|---|
| Emot | N | 1 | 2 |
| Neutral | 16 | –.358 | |
| Surprised | 16 | –.293 | |
| Sad | 16 | –.261 | |
| Happy | 16 | –.202 | |
| Angry | 16 | .378 | .378 |
| Sarcastic | 16 | | .736 |
| p = | | .165 | .273 |

**Zscore: Tempo Language = L2_Mand**

| | | Subset for alpha =.05 |
|---|---|---|
| Emot | N | 1 |
| Surprised | 16 | –.287 |
| Happy | 16 | –.242 |
| Neutral | 16 | –.144 |
| Angry | 16 | .107 |
| Sarcastic | 16 | .253 |
| Sad | 16 | .313 |
| p = | | .517 |

**Zscore: Tempo Language = Mandarin**

| | | Subset for alpha =.05 | | |
|---|---|---|---|---|
| Emot | N | 1 | 2 | 3 |
| Surprised | 16 | –.711 | | |
| Happy | 16 | –.512 | | |
| Neutral | 16 | –.214 | –.214 | |
| Angry | 16 | .013 | .013 | |
| Sarcastic | 16 | | .399 | |
| Sad | 16 | | | 1.025 |
| p = | | .064 | .089 | 1.000 |

**Appendix 2.2 Subgroups differentiated by F0_mean.**

**Zscore: F0_mean Language = Dutch**

| | | Subset for alpha = .05 | |
|---|---|---|---|
| Emot | N | 1 | 2 |
| Neutral | 16 | –.699 | |
| Surprised | 16 | –.513 | |
| Sad | 16 | –.502 | |
| Happy | 16 | –.050 | |
| Angry | 16 | | .856 |
| Sarcastic | 16 | | .909 |
| p = | | .077 | .841 |

**Zscore: F0_mean Language = L2_Mand**

| | | Subset for alpha = .05 | | | |
|---|---|---|---|---|---|
| Emot | N | 1 | 2 | 3 | 4 |
| Surprised | 16 | –.750 | | | |
| Happy | 16 | –.520 | –.520 | | |
| Neutral | 16 | –.374 | –.374 | | |
| Angry | 16 | | .122 | .122 | |
| Sarcastic | 16 | | | .473 | |
| Sad | 16 | | | | 1.050 |
| p = | | .366 | .058 | .207 | 1.000 |

**Zscore: F0_mean Language = Mandarin**

| | | Subset for alpha = .05 | | | | |
|---|---|---|---|---|---|---|
| Emot | N | 1 | 2 | 3 | 4 | 5 |
| Surprised | 16 | –1.20 | | | | |
| Happy | 16 | | –.811 | | | |
| Neutral | 16 | | | –.184 | | |
| Angry | 16 | | | | .356 | |
| Sarcastic | 16 | | | | | .769 |
| Sad | 16 | | | | | 1.070 |
| p = | | 1.00 | 1.00 | 1.00 | 1.00 | .123 |

**Appendix 2.3 Subgroups differentiated by SD_F0.**

**Zscore: SD_F0 Language = Dutch**

| | | Subset for alpha = .05 | | |
|---|---|---|---|---|
| Emot | N | 1 | 2 | 3 |
| Neutral | 16 | –.815 | | |
| Surprised | 16 | –.772 | | |
| Sad | 16 | | –.206 | |
| Happy | 16 | | .021 | |
| Angry | 16 | | | .714 |
| Sarcastic | 16 | | | 1.06 |
| p = | | .862 | .369 | .172 |

**Zscore: SD_F0 Language = L2_Mand**

| | | Subset for alpha = .05 | | | |
|---|---|---|---|---|---|
| Emot | N | 1 | 2 | 3 | 4 |
| Surprised | 16 | –.985 | | | |
| Happy | 16 | –.468 | –.468 | | |
| Neutral | 16 | | –.159 | –.159 | |
| Angry | 16 | | | .243 | .243 |
| Sarcastic | 16 | | | | .636 |
| Sad | 16 | | | | .734 |
| p = | | .069 | .273 | .156 | .192 |

**Zscore: SD_F0 Language = Mandarin**

| | | Subset for alpha = .05 | | |
|---|---|---|---|---|
| Emot | N | 1 | 2 | 3 |
| Surprised | 16 | –.887 | | |
| Happy | 16 | –.444 | –.444 | |
| Neutral | 16 | | .132 | .132 |
| Angry | 16 | | .170 | .170 |
| Sarcastic | 16 | | .276 | .276 |
| Sad | 16 | | | .754 |
| p = | | .144 | .086 | .171 |

**Appendix 2.4 Subgroups differentiated by slope of F0.**

| **Zscore: Slope of F0 Language = Dutch** | | | | |
|---|---|---|---|---|
| | | Subset for alpha = .05 | | |
| Emot | N | 1 | | |
| Neutral | 16 | –.348 | | |
| Surprised | 16 | –.335 | | |
| Sad | 16 | –.155 | | |
| Happy | 16 | –.027 | | |
| Angry | 16 | .276 | | |
| Sarcastic | 16 | .632 | | |
| p = | | .061 | | |
| **Zscore: Slope of F0 Language = L2_Mand** | | | | |
| | | Subset for alpha = .05 | | |
| Emot | N | 1 | | |
| Surprised | 16 | –.564 | | |
| Happy | 16 | –.089 | | |
| Neutral | 16 | .115 | | |
| Angry | 16 | .161 | | |
| Sarcastic | 16 | .192 | | |
| Sad | 16 | .219 | | |
| p = | | .296 | | |
| **Zscore: Slope of F0 Language = Mandarin** | | | | |
| | | Subset for alpha = .05 | | |
| Emot | N | 1 | 2 | 3 |
| Surprised | 16 | –.962 | | |
| Happy | 16 | | –.251 | |
| Neutral | 16 | | –.142 | –.142 |
| Angry | 16 | | .257 | .257 |
| Sarcastic | 16 | | .462 | .462 |
| Sad | 16 | | | .646 |
| p = | | 1.000 | .104 | .059 |

**Appendix 2.5 Subgroups differentiated by compactness.**

| **Zscore: Compactness Language = Dutch** | | | |
|---|---|---|---|
| | | Subset for alpha = .05 | |
| Emot | N | 1 | 2 |
| Neutral | 16 | –.535 | |
| Surprised | 16 | –.288 | –.288 |
| Sad | 16 | –.045 | –.045 |
| Happy | 16 | –.018 | –.018 |
| Angry | 16 | | .412 |
| Sarcastic | 16 | | .474 |
| p = | | .411 | .158 |

| **Zscore: Compactness Language = L2_Mand** | | | |
|---|---|---|---|
| | | Subset for alpha = .05 | |
| Emot | N | 1 | 2 |
| Surprised | 16 | –.789 | |
| Happy | 16 | –.203 | –.203 |
| Neutral | 16 | | .080 |
| Angry | 16 | | .168 |
| Sarcastic | 16 | | .260 |
| Sad | 16 | | .485 |
| p = | | .075 | .222 |

| **Zscore: Compactness Language = Mandarin** | | | | |
|---|---|---|---|---|
| | | Subset for alpha = .05 | | |
| Emot | N | 1 | 2 | 3 |
| Surprised | 16 | –.721 | | |
| Happy | 16 | –.453 | –.453 | |
| Neutral | 16 | | .122 | .122 |
| Angry | 16 | | .252 | .252 |
| Sarcastic | 16 | | .310 | .310 |
| Sad | 16 | | | .489 |
| p = | | .405 | .088 | .663 |

**Appendix 2.6 Subgroups differentiated by SD_intensity.**

**Zscore: SD_intensity Language = Dutch**

| | | Subset for alpha = .05 | |
|---|---|---|---|
| Emot | N | 1 | 2 |
| Neutral | 16 | –.620 | |
| Surprised | 16 | –.247 | |
| Sad | 16 | –.168 | |
| Happy | 16 | .119 | .119 |
| Angry | 16 | .185 | .185 |
| Sarcastic | 16 | | .731 |
| p = | | .101 | .146 |

**Zscore: SD_intensity Language = L2_Mand**

| | | Subset for alpha = .05 | |
|---|---|---|---|
| Emot | N | 1 | 2 |
| Surprised | 16 | –.571 | |
| Happy | 16 | –.212 | –.212 |
| Neutral | 16 | –.193 | –.193 |
| Angry | 16 | –.086 | –.086 |
| Sarcastic | 16 | | .501 |
| Sad | 16 | | .561 |
| p = | | .449 | .132 |

**Zscore: SD_intensity Language = Mandarin**

| | | Subset for alpha = .05 |
|---|---|---|
| Emot | N | 1 |
| Surprised | 16 | –.357 |
| Happy | 16 | –.046 |
| Neutral | 16 | –.036 |
| Angry | 16 | .058 |
| Sarcastic | 16 | .143 |
| Sad | 16 | .239 |
| p = | | .536 |

**Appendix 2.7 Subgroups differentiated by jitter.**

**Zscore: Jitter Language = Dutch**

| | | Subset for alpha = .05 | |
|---|---|---|---|
| Emot | N | 1 | 2 |
| Neutral | 16 | –.493 | |
| Surprised | 16 | –.210 | |
| Sad | 16 | –.184 | |
| Happy | 16 | –.129 | |
| Angry | 16 | .113 | |
| Sarcastic | 16 | | .902 |
| p = | | .325 | 1.00 |

**Zscore: Jitter Language = L2_Mand**

| | | Subset for alpha = .05 |
|---|---|---|
| Emot | N | 1 |
| Surprised | 16 | –.367 |
| Happy | 16 | –.193 |
| Neutral | 16 | .028 |
| Angry | 16 | .079 |
| Sarcastic | 16 | .125 |
| Sad | 16 | .328 |
| p = | | .351 |

**Zscore: Jitter Language = Mandarin**

| | | Subset for alpha = .05 | | |
|---|---|---|---|---|
| Emot | N | 1 | 2 | 3 |
| Surprised | 16 | –.768 | | |
| Happy | 16 | –.204 | –.204 | |
| Neutral | 16 | –.181 | –.181 | |
| Angry | 16 | | .102 | |
| Sarcastic | 16 | | .196 | |
| Sad | 16 | | | .854 |
| p = | | .146 | .570 | 1.000 |

**Appendix 2.8 Subgroups differentiated by HNR.**

**Zscore: HNR Language = Dutch**

| | | Subset for alpha = .05 | |
|---|---|---|---|
| Emot | N | 1 | 2 |
| Neutral | 16 | –.475 | |
| Surprised | 16 | –.372 | |
| Sad | 16 | –.200 | |
| Happy | 16 | .084 | .084 |
| Angry | 16 | .303 | .303 |
| Sarcastic | 16 | | .659 |
| p = | | .130 | .190 |

**Zscore: HNR Language = L2_Mand**

| | | Subset for alpha = .05 | | |
|---|---|---|---|---|
| Emot | N | 1 | 2 | 3 |
| Surprised | 16 | –.869 | | |
| Happy | 16 | –.271 | –.271 | |
| Neutral | 16 | | –.120 | –.120 |
| Angry | 16 | | .187 | .187 |
| Sarcastic | 16 | | .455 | .455 |
| Sad | 16 | | | .619 |
| p = | | .056 | .093 | .085 |

**Zscore: HNR Language = Mandarin**

| | | Subset for alpha = .05 | | |
|---|---|---|---|---|
| Emot | N | 1 | 2 | 3 |
| Surprised | 16 | –.539 | | |
| Happy | 16 | –.488 | | |
| Neutral | 16 | –.414 | | |
| Angry | 16 | .073 | .073 | |
| Sarcastic | 16 | | .464 | .464 |
| Sad | 16 | | | .904 |
| p = | | .177 | .194 | .144 |

# Curriculum Vitae

Yinyin Zhu was born on January 3rd, 1981, in Beijing, the People's Republic of China. In 2000, she was admitted to the Beijing Language and Culture University as a bachelor student, majoring in Teaching Chinese as a Second Language. She started teaching Chinese as a second language part-timely during her college, working in the international training department of Beijing Foreign Enterprise Human Resources Service Co. Ltd. After her graduation in 2004, she was employed as a full-time Chinese instructor by the same company, being responsible for teaching Chinese to foreign employees from multinational companies and embassies in Beijing, Chinese-English interpretation and Chinese culture lecture organization. From 2006 to 2008, she was pursuing her research master (M.Phil) at Leiden University in the Netherlands, specializing in phonetics and second language acquisition. In 2010, she worked as a Chinese language instructor in the Chinese Department of the Leiden University Institute for Area Studies (LIAS). In September of the same year, she was employed by the Faculty of Oriental Studies, University of Oxford, as a Chinese language instructor. During her employment in Oxford (October, 2010), she was accepted as a PhD candidate by Leiden University Centre for Linguistics, working on the perception and production of Chinese emotional prosody by Dutch L2 speakers of Chinese (long-distance). In July 2011, she came back to Leiden University to continue her PhD research on a full-time basis. This dissertation is the results of this research.