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

Zhu, Y.

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.