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

Zhu, Y.

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 95th and 5th 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.

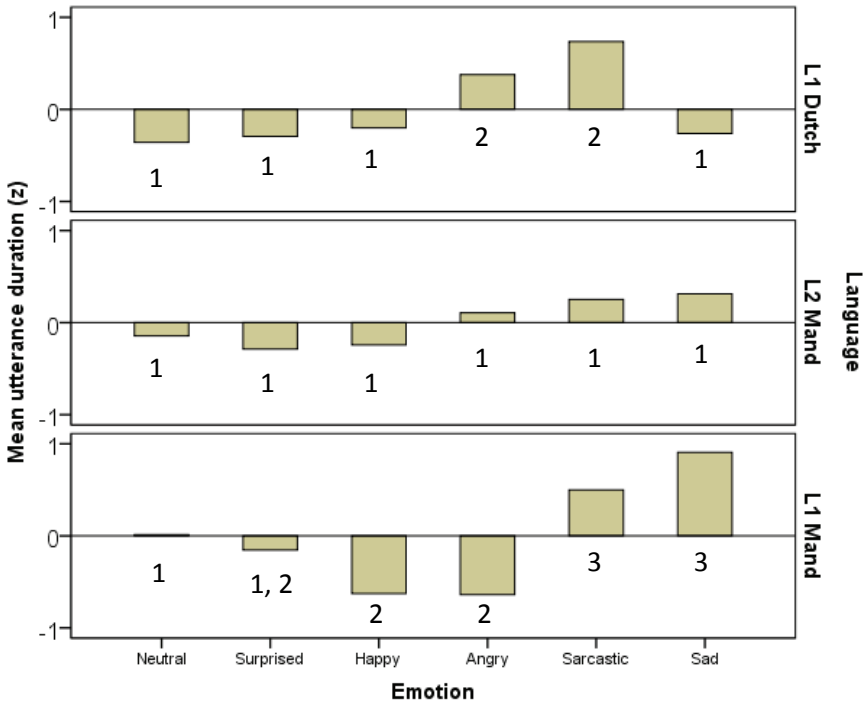

Figure 6.1. *Mean utterance duration (z-normalized within speakers) of stimuli across six emotions, classified by speaker type. 'L1 Mand' = Mandarin spoken by Chinese native speakers; 'L2 Mand' = Mandarin spoken by Dutch L2 speakers; 'L1 Dutch' = Dutch spoken by Dutch L2 speakers of Chinese. L1 Dutch 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).

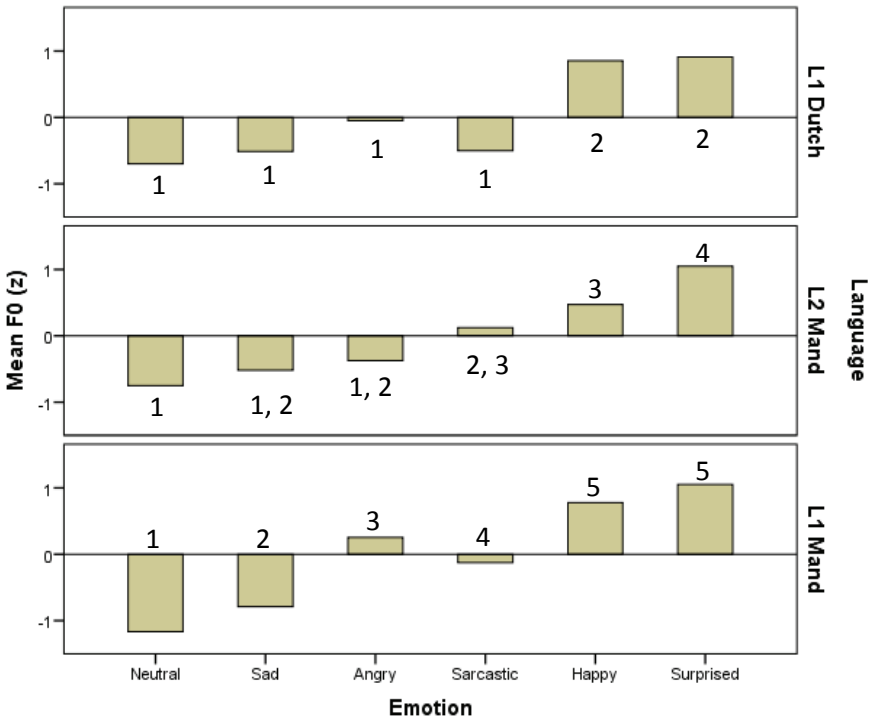

Figure 6.2. *Mean F0 (z-normalized within speakers) for six emotions broken down by speaker group (further see Figure 6.1).*

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).

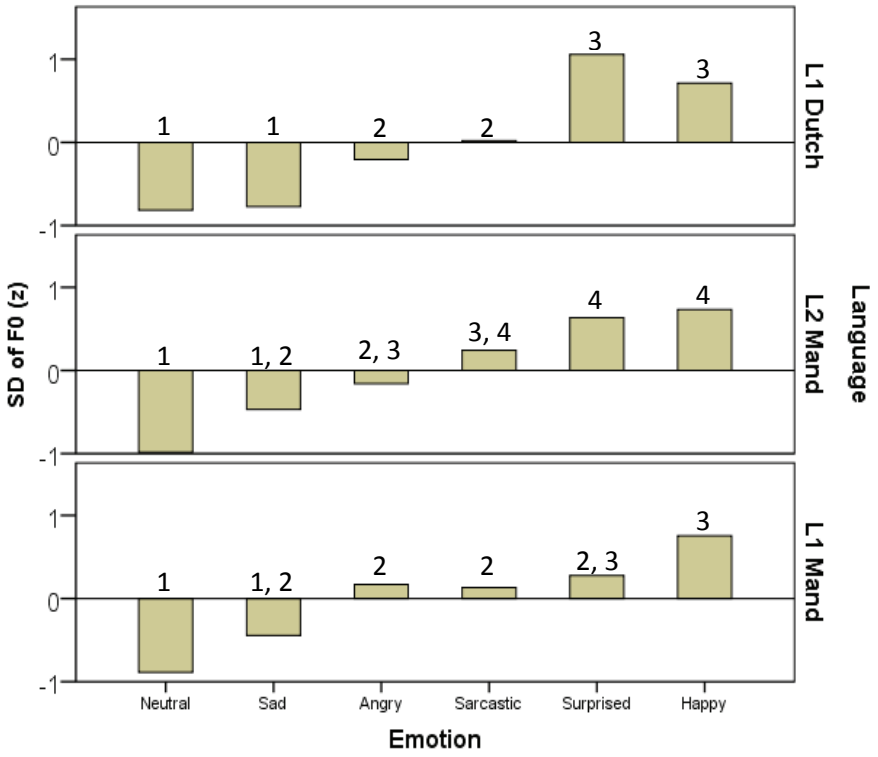

Figure 6.3. *Standard deviation of F0 (z-normalized within speakers) for six emotions broken down by speaker group (further see Figure 6.1).*

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.

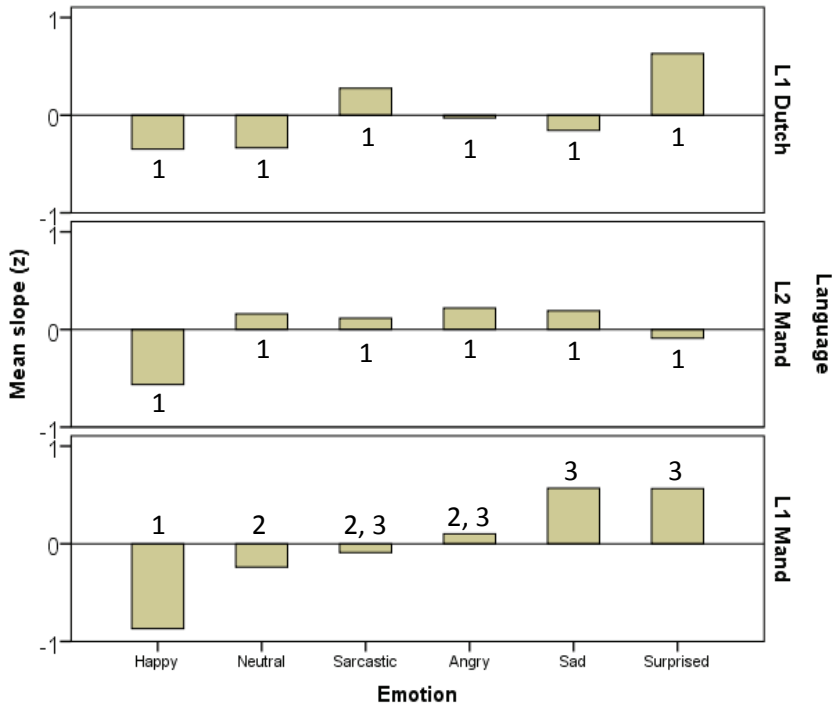

Figure 6.4. *Mean F0 slope (z-normalized within speakers) for six emotions broken down by speaker group (further see Figure 6.1).*

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.

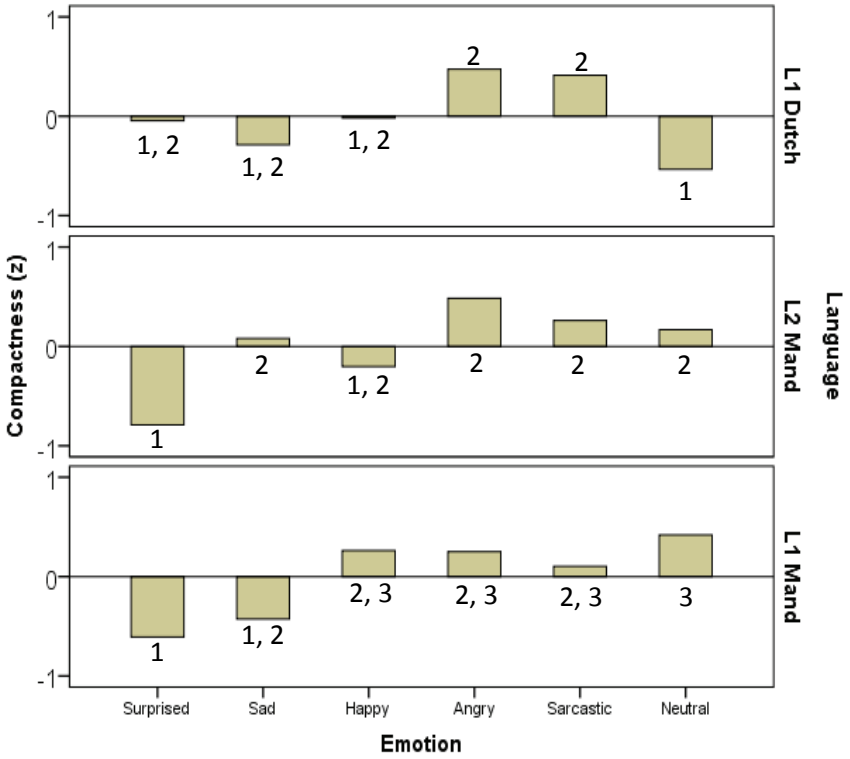

Figure 6.5. *Mean compactness (z-normalized within speakers) across the six emotions, classified by speaker group (further see Figure 6.1).*

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.

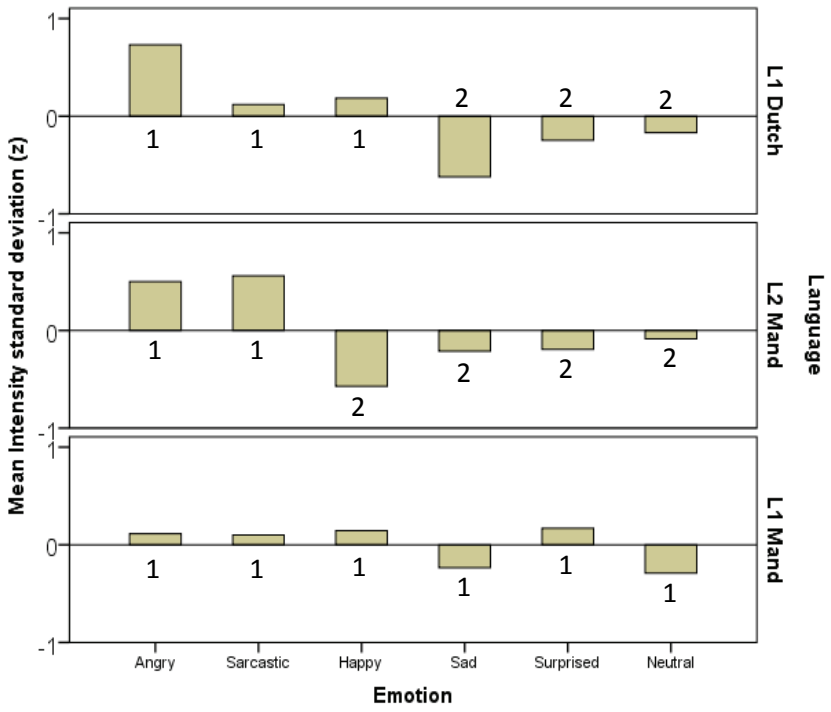

Figure 6.6. *Mean standard deviation of intensity (z-normalized within speakers) across the six emotions sorted by speaker group (further see Figure 6.1).*

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).

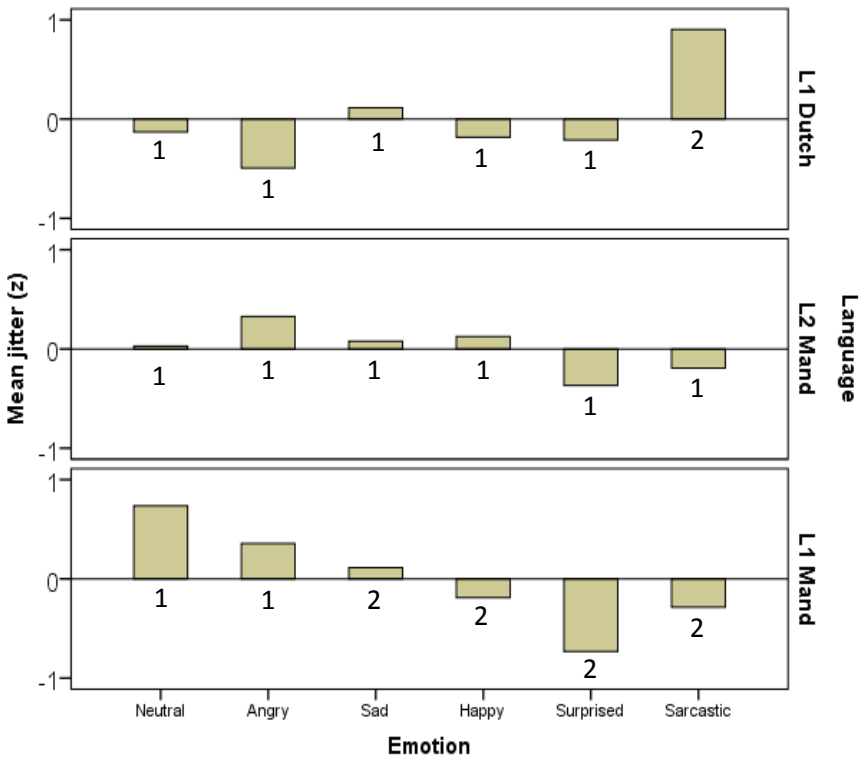

Figure 6.7. *Mean jitter (z-normalized within speakers) across the six emotions sorted by speaker group (further see Figure 6.1).*

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.).

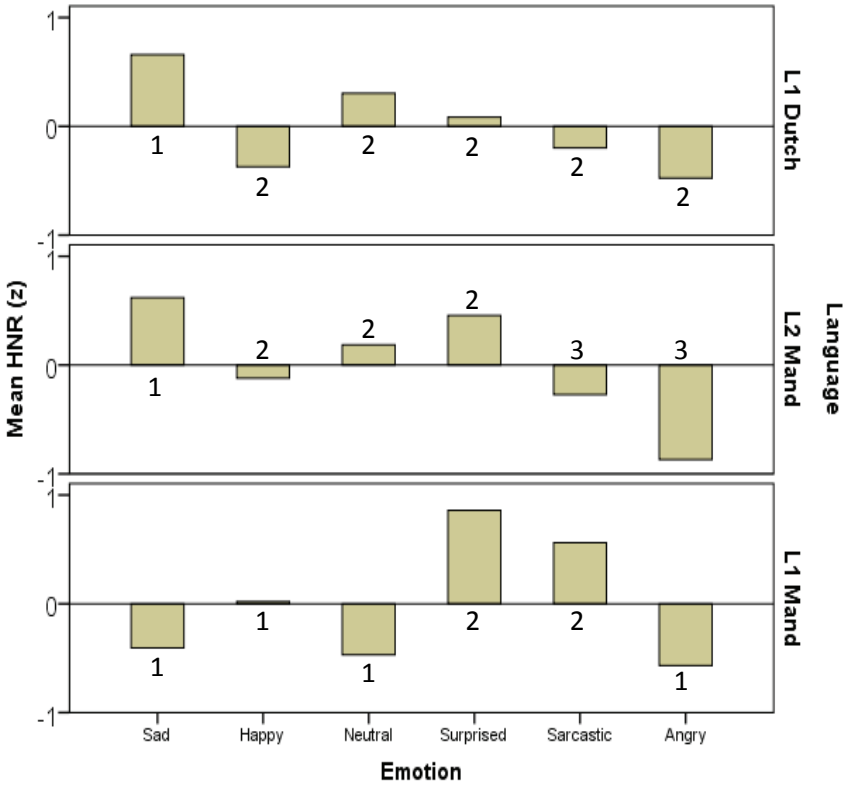

Figure 6.8. *Mean HNR (z-normalized within speakers) across the six emotions sorted by speaker group (further see Figure 6.1).*

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.