CARLOS BUSSO / MURTAZA BULUT / SUNGBOK LEE / SHRIKANTH NARAYANAN

Fundamental Frequency Analysis for Speech Emotion Processing

Let's not forget that the little emotions are the great captains of our lives and we obey them without realizing it (Vincent Van Gogh 1889)

1. Introduction

Experiencing and expressing emotions are the fundamental characteristics of human beings. Conveyed through different modalities, such as speech, facial expressions, and body language, emotions play a crucial role in decision making, action taking, and interactions for the self and for the others. Identifying and characterizing various emotions continue to be a challenging research topic for many disciplines, including but not limited to psychology, sociology, philosophy and engineering (Cowie and Cornelius, 2003) (Ekman et al., 1987) (Izard, 1977) (Picard, 1997) (Scherer, 2003).

Recent findings suggest that rational and intelligent communication between humans is closely related to emotions. Therefore, it is essential that emotions be integrated into human-machine interface (HCI) designs so that they will be more in tune with the users' emotional state and attitudes. Fortunately, recent advances in technologies such as speech recognition and computer vision are making HCIs that accommodate natural human communication modalities increasingly possible. As a result, it has added one more reason to study and model how emotions modulate and enhance the verbal and non-verbal channels of human communication.

Representation of emotions in speech in terms of measurable signal properties has been a central theme in the literature. It has been

found that speech prosody is one of the most important communicative aspects that is susceptible to emotional modulation. The intonation, tone, timing, and energy of speech are jointly modified in non-trivial manner to add emotion to the spoken utterance. Now it is well established that emotion expression can be associated with changes in prosodic and spectral characteristics of speech signals. The focus has been mainly on the statistical evaluation of pitch, duration and energy with significant emphasis on the F_0. These parameters are usually treated to be independent of each other and their relation to the emotional content of speech is described using terms showing the directional change (i.e., increase or decrease) in their value. Some of the previous research results are summarized in Table 1, which is adapted from Cowie *et al.* (2001).

The focus of this chapter is to study the effect of emotions on fundamental frequency (F_0) characteristics, as well as the effect of F_0 changes on emotion perception. Specifically, it presents results of a detailed analysis on the F_0 contour properties for understanding the emotion dependent pitch modulation observed in expressive speech. The purpose is to identify how F_0 characteristics are manipulated in expressive natural speech and to show how they can be used in the implementation of emotion recognition and synthesis systems. The results present a useful reference for the design of human-machine interaction interfaces that better incorporate expressive content. The content of this chapter is based on work drawn from our previous studies, especially those reported in Bulut and Narayanan (2008), and Busso *et al.* (2008).

	Anger	Happiness	Sadness
Pitch Statistics	Higher in mean, median, range and variability	Higher in mean, median, range and variability	Lower in mean and range
Pitch Changes	Abrupt on stressed	Smooth, upward inflections	Downward inflections
Pitch Contour	Angular frequency curve, stressed syllables ascend frequently and rhythmically, irregular up and down inflection, level average pitch except for jumps of about a musical fourth or fifth on stressed syllables	Descending line, melody ascending frequently and at irregular intervals	Downward inflections

Table 1. Relations between F_0 and emotions. Neutral speech is considered as a reference for the presented relative descriptions (from Cowie *et al.* (2001)).

The present chapter is organized as follows. In Section 2, fundamental frequency analysis methods and results reported in the literature are summarized. Also, our approach to compare the F_0 properties of emotional and neutral speech are presented (Bulut *et al.*, 2007) (Busso *et al.* , 2008) (Yildirim *et al.* , 2004). In the next section (Section 3), pitch features are investigated in terms of their usefulness for emotion recognition, by analyzing their performance in the binary emotion classification tasks between neutral and emotional speech (Busso *et al.*, 2008). Next, in Section 4, the F_0 characteristics of emotional utterances are modified, and their effect on emotional content and speech quality perception is investigated (Bulut and Narayanan, 2008). Finally, discussion and conclusions are presented in Section 5.

2. Analysis of fundamental frequency

It is well accepted that F_0 contour conveys emotional information, especially for certain emotional categories. For instance, many studies have found that recorded happy or angry utterances have relatively higher and more variable pitch level accompanied by a wider pitch range. In contrast, for subdued emotional categories such as sadness, the pitch average is lower relative to neutral speech (Banse and Scherer, 1996) (Juslin and Laukka, 2003) (Murray and Arnott, 1993) (Scherer, 1986). This section discusses previous efforts toward identifying the emotionally salient aspects from speech and presents the results of our analyses.

Lieberman and Michaels (1962) analyzed whether perturbations of fundamental frequency convey emotional information. They performed subjective experiments, in which the participants were asked to assess the emotional content of semantically neutral sentences recorded over different emotions. Using an analysis by synthesis approach, they manipulated the pitch, and the intensity values; for example, smoothing the F_0 contour using 40-msec and 100-msec windows. Combinations of these modifications were presented to the labelers. The results indicated that the identification of the emotional content decreased when the F_0 contour is smoothed, suggesting that small localized perturbations are indeed important in the perception of emotions. Similar perceptual experiments were presented by Breitenstein et al. (1986). Although their results showed that small pitch variations were not the most important cues in the perception of emotion, they did influence the emotional percepts.

Paeschke et al. (1999) studied several features, derived from the F_0 contour, that could be used to characterize expressive speech. Their results suggested that the shape of the fundamental frequency contour conveys emotional information. They found that the slope, and the steepness of rising and falling of F_0 present different patterns for the targeted emotional categories (anger, fear, happiness, boredom, disgust and sadness).

In several languages, the values of the F_0 contour gradually decrease at the end of the sentences, a phenomenon known as pitch declination. Wang et al. (2005) analyzed whether the patterns in pitch declination is affected by emotional modulation. They focused on contrasting happy and neutral sentences, using 4-word sentences in Mandarin. They concluded that the declination degree was lower for happy sentences, especially in the last part of the sentence. These results agree with our previous study, in which we showed that the first and second part of a sentence present different F_0 patterns, which are also emotionally dependent (Fig. 1) (Bulut et al., 2005). The study presented by Paeschke (2004) also indicates that declination is an important emotional cue, especially for emotional categories such as boredom and sadness.

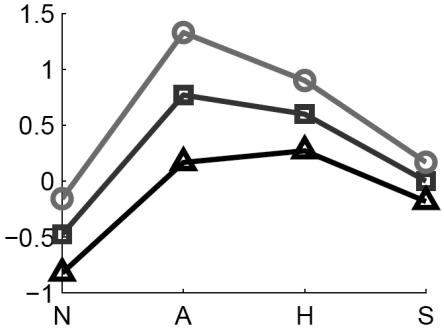


Fig. 1. Main effects of emotion and position in the sentence for F_0 median. Results for the first and second part of the sentence are marked with 0 and Ω, respectively. The values over the entire sentence are marked with . Symbols N, A, H and S represent neutral, anger, happiness and sadness (from Bulut *et al.* (2005)).

Although the aforementioned aspects of the F_0 contour convey emotional information, some studies have suggested that gross statistics from pitch such as mean and range conveyed most of the emotional information (Bänziger and Scherer, 2005) (Ladd *et al.*, 1985). For example, the analysis presented by Bänziger and Scherer (2005) indicates that mean levels and the range of the F_0 contour account for most of the important emotional variation found in the pitch. In contrast, the shape of the F_0 contour was only slightly affected by emotional modulation. Scherer *et al.* (1984) suggested that different aspects of the pitch are associated with linguistic and

paralinguistic communicative goals. They suggested that the shape of the F_0 contour is associated with the grammatical structure of the sentence. Therefore, the interplay between affective and linguistic goals in the pitch shape is tighter compared to the degree of freedom observed in the manipulation of gross patterns in the F_0 contour.

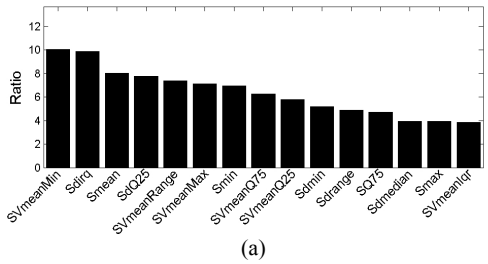
Our recent experiments were designed to identify the most emotionally salient features of fundamental frequency (Busso et al., 2008). The approach was based on contrasting the F_0 features derived from neutral and expressive speech. The data support the idea that the gross statistics from the pitch are the most emotionally prominent aspects of the F_0 contour. The remainder of this section will discuss the results from this work.

In the study, sixty pitch features were extracted from expressive speech and compared to their neutral counterpart extracted from a reference neutral corpus (WSJ1 (Paul and Baker, 1992)). There were 39 features corresponding to global sentence-level statistics (one set of features from each utterance), and 21 features corresponding to local statistics computed at voiced-level regions, defined as the speech segments with consecutive voiced frames (one set of features for each voiced region in the utterance). The pitch features included the most common statistics used for emotion recognition such as the mean, range, minimum, and maximum of the pitch contour. They also contained information about the F_0 shape, such as the inflection, curvature and slope, which were parameterized by fitting a first, second and third order polynomial, respectively. In the analysis, three emotional corpora were used: the electromagnetic articulography (EMA) database (Lee et al., 2005), the German emotional speech (GES) database (Burkhardt et al., 2005), and the emotional prosody speech and transcripts (EPSAT) database (Liberman et al., 2002). These corpora that were used in the original work include different speakers, emotional categories, and also different languages (English and German). In this chapter, we limit the analysis to English speech. Therefore, the results reported here and in Section 3 were recomputed without the GES database.

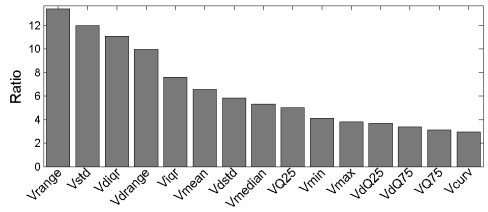
Instead of comparing the mean values of the expressive pitch features and their neutral counterpart, we proposed to compare their distributions by using Kullback-Leibler Divergence (KLD) (Cover and Thomas, 2006). For a given emotional corpus, we estimated the

ratio between the KLD from its emotional set and the KLD from its neutral set. A high value of this ratio would indicate that the distribution of this pitch feature changes during expressive speech. This allows us to rank the pitch features according to their emotional prominence, when compared to neutral speech.

Figures 2(a) and 2(b) provide the most salient pitch features in terms of the average KLD ratio across emotional categories and databases, for sentence- and voiced-level features, respectively. The figures show that mean (median), range (interquartile range), maximum (upper quartile) and minimum (lower quartile) are the most emotionally prominent pitch features. The average inflection, curvature and slope are not included among the top 15 most prominent sentence-level features. This result indicates that the F_0 shape is not as emotionally salient as gross pitch statistics. As suggested by Scherer et al. (1984), the linguistic aspects of the utterance may have stronger influence in the resulting pitch shape. Further discussion is given in Section 5.



(a)



(b)

Fig. 2. Most emotionally prominent features according to the average KLD ratio between features estimated from expressive and neutral speech. (a) sentence-level features, (b) voiced-level features (from Busso *et al.* (2008)).

The results from this analysis also indicate that variations within voice segment regions are emotionally salient. Figure 2(b) shows that the patterns of the standard deviation in voiced regions, *Vstd*, changes during expressive speech. This result agrees with the work presented by Lieberman and Michaels (1962). While these variations are useful for emotion recognition (Section 3), they do not significantly change the (human) emotional perception of the sentence, as discussed in Section 4.

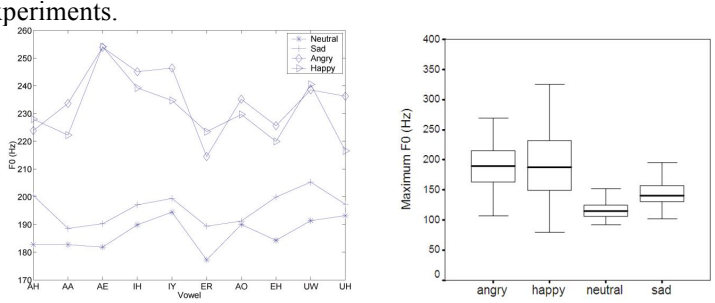
3. Automatic emotion recognition from F_0 features

Many studies have proposed the use of features derived from the fundamental frequency contour as one of the key information sources for automatic emotion recognition (Pantic and Rothkrantz, 2003). A common approach is to extract as many derived features as possible, and then use feature selection techniques to find a reduced subset that maximizes the performance (Schuller *et al.*, 2007). Some of the most common selected F_0 features are the mean, range, minimum, maximum and standard deviation statistics from the F_0 utterance contour (Clavel *et al.*, 2008).

In previous work, we have analyzed the F_0 patterns at different segmental and linguistic levels (phoneme, word, part of speech) (Bulut et al., 2005) (Lee et al., 2005) (Yildirim et al., 2004). For example, we observed that F_0 mean significantly differs for angry, happy, sad and neutral speech (Fig. 3). Using F_0 features, the recognition rate for this four-class problem was between 50.9% and 55.7% (chance level is 25%) (Lee et al., 2004) (Yildirim et al., 2004). Most of the errors in the classifier came from the confusion between happiness and anger, and sadness and neutral state. These emotional categories are similar in the activation domain (calm-excited), but they differ in the valence (positive-negative) domain (Fig. 6).

Bänziger and Scherer (2005) suggested that the fundamental frequency is mainly affected by the arousal level of the utterance. They analyzed changes in the F_0 contour in terms of the degree of activation in the sentences. They also analyzed the change of F_0

contour in terms of emotional categories. They found that changes in the arousal level change fundamental frequency. However, they did not find evidence for the qualitative changes in the F_0 contour among different emotions. These observations agree with the confusion between emotional categories observed in our aforementioned experiments.



(a) (b)

Fig. 3. Emotional differences observed in F_0. (a) Pitch mean in terms of vowels (from Yildirim *et al.* (2004)). (b) Maximum pitch distribution (from Lee *et al.* (2005)).

An alternative approach to identify useful pitch features for emotion recognition is to perform binary classification between each individual emotional category and neutral speech (e.g., neutral-happiness, neutral-anger). This approach also provides insights into the emotional categories that can be accurately recognized from F_0 patterns. In our previous work, logistic regression models were proposed for this binary classification problem (Busso et al., 2008). A nice property of this framework is that the significance of the contribution of the features in the model can be quantified with the use of log-likelihood ratio tests.

Figures 4(a) and 4(b) show the pitch features, which on average have the highest impact in terms of log-likelihood, when individually included in the logistic regression models (one feature at a time). The top sentence-level features are the median (Smedian), and mean (Smean). The top voiced-level features are the upper quartile (VQ75), median (Vmedian), and mean (Vmean). The features depicting the pitch shape were not among the top features according to this criterion. These results support the idea that gross

pitch statistics are the most emotionally salient aspect of the F_0 contour.

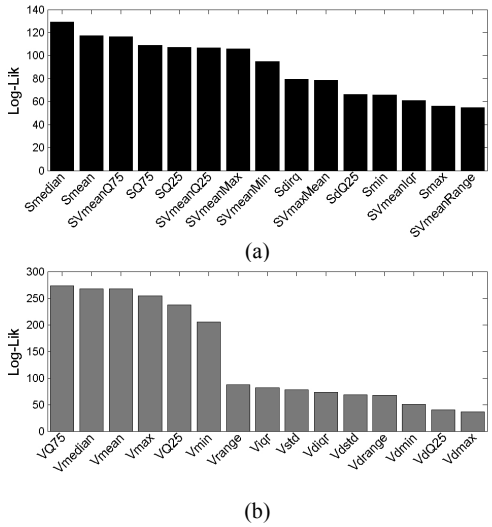


Fig. 4. Most emotionally prominent features according to the average log-likelihood score in logistic regression analysis, when a single feature is entered at a time in the model. (a) sentence-level features, (b) voiced-level features (from Busso *et al.* (2008)).

In Figures 4(a) and 4(b), the logistic regression models contain only a single feature. Therefore, the best features under this criterion may provide redundant information or be highly correlated. A second experiment was also proposed in this study to address this question. Instead of including one feature at a time, *Sequential Forward Feature* (SFF) selection was used to add pitch features until the improvement in the model was not significant. This procedure was performed for each emotional category in each emotional database (EMA and EPSAT). Figures 5(a) and 5(b) show the pitch features that were most frequently selected. Under this criterion, the top features are the median (Smedian), and the derivative of the median (Sdmedian) for the sentence level features, and the minimum (Vmin), and curvature (Vcurv) for the voiced-segment features.

Although Vcurv and SVmaxCurv are not among the most prominent emotional features (Figs. 2 and 4), the results indicate that the pitch curvature provides complementary information that is important for emotion recognition purposes. Notice that the pitch mean value at sentence-level (Smean), which was the second best feature when a single feature was included in the model (Fig. 4(a)), does not appear in Figure 5(a). Given that the mean and median are highly correlated ($\rho \approx 0.96$), it is not surprising that these two features were almost never selected together.

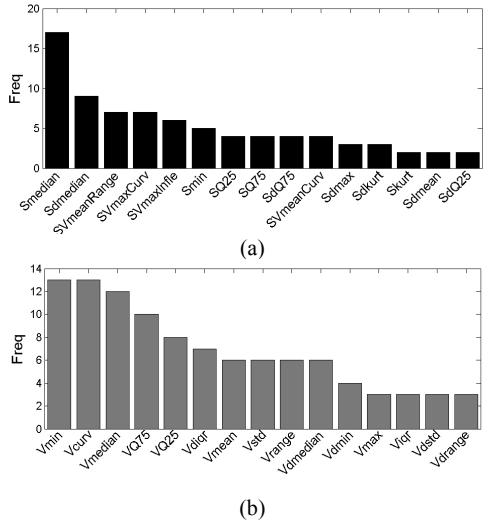


Fig. 5. Most frequently selected features in logistic regression models, when the features are added using sequential feature selection. (a) sentence-level features, (b) voiced-level features (from Busso *et al.* (2008)).

Table 2 gives details of the logistic regression models built with sequential forward feature selection. The results are separately presented for each emotional category and database considered. The results indicate that certain emotional categories such as sadness and boredom cannot be discriminated from neutral speech using the investigated pitch features.

An interesting observation is that emotional categories with similar activation level cannot be easily recognized from each other using the proposed F_0 features. Table 2 shows that the performance of the logistic regression models for the emotional categories with neutral activation level is low (Nagelkerke R2 < 0.5). For better visualization, Figure 6 presents the approximate location of different emotional categories in the activation-valence space. The locations determined were based on the Feeltrace snapshots presented by Cowie and Cornelius (2003) and Cowie et al. (2001). The emotional categories in which the Nagelkerke r-squares of the logistic regression models were higher than 0.5 (Table 2) were highlighted in black (otherwise they were plotted in gray). This figure suggests that pitch contour is mainly modified according to the arousal level. This observation agrees with the conclusion made by Bänziger and Scherer (2005). It also explains why emotional categories such as happiness and anger, which are different in the valence domain, are usually confused when only pitch features are used (Yildirim et al., 2004). As suggested by Ladd et al. (1985), voiced quality features may provide information to discriminate in the valence domain.

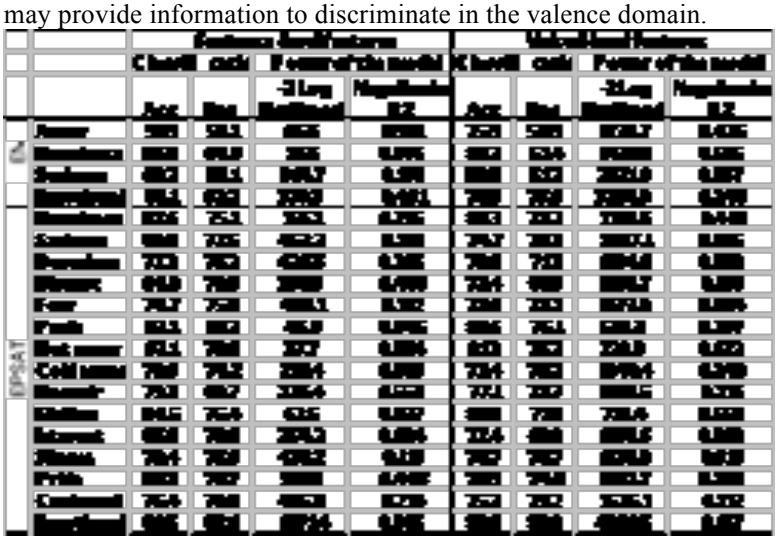


Table 2. Details of the logistic regression models using SFF (from Busso *et al.* (2008).

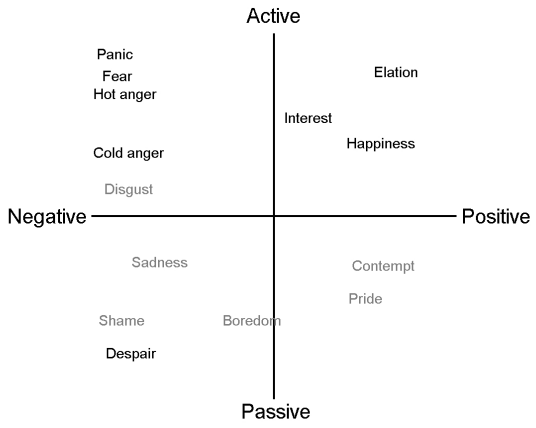


Fig. 6. Location of the emotional categories in the activation-valence space. For the emotional categories in gray, the power of the logistic regression model was inadequate to accurately recognize emotional from neutral speech ($r^2 < 0.5$). The figure was adapted from Cowie and Cornelius (2003) and Cowie *et al.* (2001).

4. Analysis by synthesis: emotional effects of F_0 feature modifications

Section 2 suggests that the gross statistics from the fundamental frequency contour are the most emotionally prominent F_0 features. However, the results presented in Section 3 indicate that F_0 features that were not among the most emotionally prominent features (i.e., F_0 curvature) were also useful for other applications such as automatic emotion recognition. In this section, we address the following question: What aspects of the F_0 contour have stronger influence on the human perception of expressive speech?

In the literature, the synthesis of emotional speech has been studied from different perspectives using formant (Burkhardt et al., 2000) (Cahn, 1990) (Murray and Arnott, 1995) and concatenative

(Bulut et al., 2002) (Bulut et al., 2005) (Montero et al., 1999) (Murray et al., 2000) (Raux and Black, 2003) synthesis techniques. These studies provide useful information about how different prosody and spectral parameters need to be jointly modified in order to synthesize emotional speech. For example, Burkhardt and Sendlmeier (2000) showed that broader pitch range, when combined with a faster speech rate and modal or tense phonation, can increase the perception of joy. For anger, the effects of pitch modification were less obvious. Our copy synthesis experiments on English (Bulut et al., 2002) showed that prosody was more descriptive for sadness and neutral emotions, while segmental inventory was the main factor for anger. A similar approach applied for Spanish concluded that prosody affects emotions differently and that sadness and surprise were better associated with supra-segmental variations, while happiness and cold anger were more segmental (Montero et al., 1999).

The aforementioned techniques for studying emotions have a powerful value. However, they do not specifically address how important the F_0 characteristics are for the perception and synthesis of emotions. The main reason is that many different parameters are jointly modified.

The analyses presented in the previous sections indicate which F_0 features are the most emotionally prominent features. In this section, we explore whether these aspects of the F_0 contour can be individually manipulated to change the human emotional perception of an utterance. Following an experimental methodology, we systematically modified the F_0 contour characteristics of natural emotional utterances to observe their influence in the perception of the emotions (Bulut and Narayanan, 2008). This simple procedure allowed us to infer the role of F_0 properties in emotion perception. It also provided important insights into the limits within which the F_0 characteristics of the utterances can be modified without significantly altering the original emotion. The remainder of this section presents the results from this study.

To reduce the number of modifications, we limited the analysis to the mean, range and shape of the fundamental frequency contour. We proposed an analysis by synthesis approach. Two sentences, “She told me what you did” (sentence 1) and “This hat makes me look like

an aardvark" (sentence 2) were recorded by a female (speaker 1) and a male speaker (speaker 2). The sentences were uttered portraying the following emotions: anger, happiness, sadness and neutral state. In total, 16 utterances were considered (two speakers, two sentences, four emotions).

Several modifications manipulating the mean, range and shape of the natural F_0 contours were applied to all of the recorded emotional utterances. These three parameters were chosen because of their common usage in emotion analysis and recognition studies as explained in the previous sections. The modifications were done using the Time Domain Pitch Synchronous Overlap and Addition (TD-PSOLA) algorithm (Moulines and Charpentier, 1990) as implemented in the Praat software (Boersma and Weeninck, 1996).The applied modifications can be categorized into three groups: mean, range and stylization modifications (summarized in Table 3). The details are given below.

- Modifications in F_0 mean: The mean was modified by shifting the F_0 contour up or down. The following modifications were applied: (1) Increasing/decreasing the original F_0 mean by 10%, 15%, 25% and 50%, (2) Making the F_0 mean equal to 50, 100, 150, 200, 250 and 300 (Hz).
- Modifications in F_0 range: The range was modified by multiplying the F_0 contour with a constant and then shifting the contour up or down so that the mean will be the same as the original mean value. The following modifications were applied: (1) Scaling the range by 0.5, 0.75, 1.5 and 2, (2) Making the F_0 range equal to 10, 30, 50, 80, 110 and 150 (Hz).
- Stylization modifications: The shape of the F_0 contour of the utterances was altered by stylizing the F_0 contour. The following modifications were applied: Stylizing the F_0 contour by a 2, 5, 10, 15, and 40 semitone frequency resolution. Stylization of the F_0 contour was performed using the Praat software. The logic behind the stylization algorithm is to try to represent the F_0 contour using linear segments. The length of the linear segments was determined by the frequency resolution component. For instance, while a 2 semitone resolution corresponds to fairly short linear segments, thus preserving the general contour shape, 40 semitone resolution may cause the whole utterance F_0 contour to be a line.

F0	Mean	Range	Stylization
Increase	m1: 0.10	r3: +50%	
	m2: 0.15	r4: +100%	
	m3: 0.25		
	m4: 0.50		
Decrease	m5: -0.10	r1: -50%	
	m6: -0.15	r2: -25%	
	m7: -0.25		
	m8: -0.50		
Value	m9: =50	r5: 10	s1: =2
	m10: =100	r6: 30	s2: =5
	m11: =150	r7: 50	s3: =10
	m12: =200	r8: 80	s4: =15
	m13: =250	r9: 110	s5: =40
	m14: =300	r10: 150	

Table 3. Summary of the performed F_0 contour modifications. The values for mean and range are in Hz and the values for stylization are in semitones.

All natural and resynthesized utterances were evaluated by listening experiments with naïve listeners that included both native and non-native American English speakers. Before evaluation, all speech files were normalized so that the maximum digitized waveform amplitude was 1. In the listening tests – conducted in a quiet room, using headphones and with a single rater at a time – first the speech file was presented and then the raters were asked to choose among the following options: Happy, angry, sad, neutral, and other. The raters were particularly instructed to choose *other* if their choice of emotion was not listed, or if they could not decide on the emotional content, or if the speech sounded to them as a mixture of several emotions. They were allowed to listen to each utterance as many times as they liked before making their decision. After the raters had chosen the emotional content label, they were asked to rate the naturalness (i.e., quality) of the utterance on a scale from 1 to 5, with five corresponding to the most natural. They were specifically instructed to give low values if the speech was perceived to be different from natural human speech in terms of quality. Again, the raters were able to listen to the speech as many times as they liked. The files were presented in a different random order for each rater.

In order to limit the time of any single test to around 20 minutes the test set was divided into 10 groups of 48 utterances, each consisting of three variations (which were chosen randomly) of the 16 original utterances. Average number of raters per set was 9.2 and

each set was evaluated by at least seven raters. In total, there were 14 different people that participated.

A summary of the listening tests results is presented in Tables 4, 5, and 6. The average values of percentage, quality, and similarity variables are displayed in the tables. The percentage variable represents the percentage of listeners that perceived the same emotion as the original emotion. The quality variable is a measure of speech quality as determined by the responses of the listeners. Five represents excellent quality, while four, three, two, and one correspond to good, fair, poor, and bad quality ratings, respectively.

	m1	m2	m3	m4	m5	m6	m7
	10%	15%	25%	50%	-10%	-15%	-25%
Similarity	0.97	0.97	0.93	0.8	0.96	0.97	0.92
Percentage	0.74	0.76	0.71	0.59	0.74	0.74	0.65
Quality	4.29	4.29	4.07	3.57	4.27	4.09	3.4
	m8	m9	m10	m11	m12	m13	m14
	-50%	=50	=100	=150	=200	=250	=300
Similarity	0.88	0.81	0.88	0.95	0.93	0.92	0.82
Percentage	0.63	0.57	0.63	0.73	0.73	0.74	0.63
Quality	2.2	1.99	2.73	3.18	3.63	3.49	3.42

Table 4. Summary of the F_0 mean modification effects.

The difference between the perception of original (unmodified) utterances, and resynthesized utterances was quantified using the similarity measure defined in Equation 1. To do that, first, the utterances were represented by normalized vectors (i.e., all entries sum to 1) consisting of fractions of each emotion as determined by human raters. For example, a vector y = [0.5 0.3 0.1 0 0.1] was used for an utterance that was perceived as happy, angry, sad, neutral and other by 50%, 30%, 10%, 0% and 10% of human raters, respectively. Then, these vectors were used to calculate the similarity value. Technically, the similarity function is a simple nonmetric measure that is the cosine of the angle between two vectors and that has a large value (i.e., close to 1) when the vectors point in the same direction (Duda *et al.*, 2000). In our case, a large similarity value indicates that the effect of modification was minimal and that the subjective recognition percentages of the original and its modified version were similar. In contrast, a low similarity value indicates that the performed modification changed the emotional content of the original utterance.

$$s(x, y) = \frac{x' y}{\|x\| \cdot \|y\|} \tag{1}$$

The results show that although F_0 modification caused new emotional nuances to be perceived, the major emotion perception was not significantly affected in many cases. In other words, even the significant variations in F_0 parameters did not mask the original emotion perception. In general, it was observed that the F_0 modifications caused sad, neutral, or other emotion perception to increase, and angry, or happy emotion perception to decrease. The effects of the F_0 range modifications on emotion perception were more prominent than the F_0 mean modifications. Also, for the F_0 range modifications, the drop in the perceived speech quality was less than the F_0 mean modifications. These results suggest that one should focus on F_0 range and not F_0 mean modifications during the synthesis of emotional speech. The results also indicate that eliminating the small prosodic variations (s1, s2, s3) in the F_0 contour shape did not significantly decrease the perception of the original emotions. It was only when the F_0 contour at the sentence level was fully linearized – eliminating any accents and foot patterns (Klabbers and Van Santen, 2004) – the percentages of emotions changed significantly. This is an important result which has implications for emotional speech synthesis. It indicates that during the emotion synthesis, priority should be given to the other modifications (such as overall F_0 contour shape, F_0 range, and spectral characteristics), but not to the small prosodic variations in the F_0 contour shape. Later, in order to improve the naturalness of the synthesized speech, small prosodic variations may be considered. As a final point, it is also important to note that the results were significantly dependent on the speaker and the original utterance characteristics. The readers are referred to (Bulut and Narayanan, 2008) for a detailed discussion of the results presented in this section.

	r1 (-50%)	r2 (-25%)	r3 (+50%)	r4 (+100%)	r5 (=10)
Similarity	0.91	0.97	0.94	0.93	0.73
Percentage	0.63	0.76	0.74	0.65	0.51
Quality	3.95	4.35	4.13	3.56	3.45
	r6 (=30)	r7 (=50)	r8 (=80)	r9 (=110)	r10 (=150)
Similarity	0.69	0.76	0.83	0.91	0.84
Percentage	0.47	0.51	0.58	0.69	0.61
Quality	3.52	3.76	3.84	3.9	4.08

Table 5. Summary of the F_0 range modification effects.

	s1 (=2)	s2 (=5)	s3 (=10)	s4 (=15)	s5 (=40)
Similarity	0.97	0.97	0.91	0.83	0.65
Percentage	0.78	0.72	0.6	0.58	0.48
Quality	4.51	4.35	3.73	3.44	3.01

Table 6. Summary of the F_0 stylization effects.

5. Discussion and conclusions

According to the results presented in this chapter, we conclude that the most prominent features are gross (overall phrase/utterance level) statistics from the F_0 contour such as the mean, minimum, maximum, and range. In contrast, the F_0 shape features were not among the most emotionally salient features. Since affective and linguistic goals are simultaneously conveyed in the pitch, some kind of interplay needs to regulate these two production processes. It seems that the shape of the fundamental frequency is constrained by the lexical content of the utterance, as suggested by Scherer *et al.* (1984). Therefore, emotional information seems to be mainly encoded by modulating the global patterns in the pitch. As an aside, a similar interplay is observed in the face, in which the upper face area is less constrained by the articulatory process, and therefore, it has more degrees of freedom to convey emotions (Busso and Narayanan, 2006).

Although this study indicates that major acoustic correlates of emotion expression are global pitch parameters such as mean and range, it is possible that these observations are the surface phenomena of the underlying dynamics associated with the shape of the pitch contour and the rate of pitch change. Therefore, further attention should focus on the dynamic aspects of pitch modulation in expressive speech.

From an automatic emotion recognition perspective, it is interesting to note that emotional categories which exhibit neutral level of activation were not accurately discriminated from neutral speech using F_0-based features. This result supports the studies that indicate that fundamental frequency is affected by the arousal level, rather than by emotional categories (Bänziger and Scherer, 2005). Therefore, F_0 features may not be adequate by themselves for performing multi-class emotional categorization. Other aspects of the speech also convey important emotional information (Lieberman and Michaels, 1962). For instance, in our previous work, the performance of an automatic emotion classifier increased from 50%, when only pitch-related features were used, to 67%, when duration and energy were included (for a similar task, human recognition was only 68.3%) (Yildirim et al., 2004). Likewise, other aspects of speech such as voice quality features may provide information to discriminate expressive speech in the valence domain, as suggested by Ladd et al. (1985).

Although F_0 shape features were not emotionally salient compared to global F_0 patterns, they do provide supplementary information that can be used to recognize expressive from neutral speech, which agrees with the analysis presented by Paeschke (1999). Even though some features are not emotionally prominent, they may still provide additional emotional cues. Notice that features that are good to discriminate expressive speech types are not necessarily useful to modify the emotional content of an utterance. For example, the pitch shape needs to be dramatically modified to significantly change the emotional content of the speech. Likewise, changes in the F_0 mean (median), which was one of the most emotional prominent features in our analysis, do not influence the emotion perception as much as the F_0 range modifications, which was always ranked below

the mean. These results indicate that relevant features for synthesis and recognition of expressive speech need to be studied separately.

From an expressive speech synthesis perspective, the results discussed here suggest that, in general, modifications in the mean, range and shape of fundamental frequency introduce new emotional nuances, and therefore change the perception of emotions. However, in order for the (emotional) change to be significantly different than the perception of emotions of the input speech, the modifications should be performed in larger scales. This result indicates that the emotion content is not affected by the small variations in F_0 mean, range, and shape values.

In light of the presented analyses, we suggest the following improvements for systems analyzing the variations of F_0 characteristics in emotional speech, following Bulut and Narayanan (2008). These suggestions can be easily generalized to other parameters as well.

- Use of more descriptive linguistic labels: A hybrid labeling scheme combining categorical and attribute descriptions can be utilized. For example, considering the findings showing that valence, activation and intensity dimensions are correlated with the acoustic features of emotional speech (Grimm et al., 2007) (Schröder et al., 2001), an angry utterance can be described as angry, high (low, medium) activation, high (low, medium) valence, high (low, medium) intensity, instead of just angry. As a recent example, this technique was used by Bänziger and Scherer (2005) to study F0 variations in emotional speech.

- Clear definitions of the labeling specifications: Each of the chosen emotional labels should be described in detail in terms of what type of emotions they represent. For example, emotional adjectives such as annoyed, hostile, impatient, intolerant, nervous, etc. can all be used to describe angry emotion. This is important because, due to the subjective nature of emotions, identical labels used by different research studies may correspond to very different emotions.

- Careful design of sentence structure for corpus collection: Concepts such as constituency, grammatical relations, and sub-categorization and dependencies (Jurafsky and Martin, 2000) can be employed to restrict the 'text space'. Sentences can be chosen to have

similar syntax and grammar, for example. In addition, the lexicon can also be restricted, because different words may have different effects as suggested by 'The dictionary of affect' (Whissel, 1989). Note that imposing such restrictions on text may be difficult to generalize, but it will be helpful for modeling the acoustical correlates of emotional speech. Once successful parameterization is achieved, the models can be adapted to more general datasets.

These suggestions are in line with the four main issues, the scope, naturalness, context of the content, and the kinds of descriptor it is appropriate to use, emphasized by Douglas-Cowie et al. (2003) for more systematic database construction and analysis. They also support the idea that acoustic parameters should be studied in connection with the human perception of prosodic and paralinguistic features (Roach, 2000).

Perception of information present in speech signals is dependent on linguistic, expressive, physiological, and perspectival quality factors (Modulation Theory (Traunmüller, 2005)). The first three of these factors have significant effects on the acoustic features of speech and therefore should be accounted for during the analysis of emotional information in speech.

The Facial Action Coding System (FACS) developed by Ekman, Friesen and their colleagues (Ekman and Friesen, 1978) (Ekman et al., 1987) consists of rules for reading and interpreting facial emotions. It is desirable to have a similar system for speech. The emotional labels used for facial expressions can be too restrictive to represent all possible variations in speech. For example, they may not be adequate to represent different levels of activation, valence, and intensity that is present in emotional speech. Therefore, more expressive emotional labels, combining dimensionality approaches can be developed. Our hope is that there will be more research on defining expressive descriptors particular to emotional speech.

Acknowledgments

This research was supported in part by funds from the National Institutes of Health (NIH), National Science Foundation (NSF), and the Army. The authors thank colleagues in the emotion research

group for their support and valuable comments. Thanks go to Emily Mower, Serdar Yildirim, Chul Min Lee, Abe Kazemzadeh, Matthew Black, Chi-Chun Lee, Angeliki Metallinou, Jeannette Chang, and Samuel Kim.

References

Banse, Rainer / Scherer, Klaus R. 1996. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70/3, 614–636.

Bänziger, Tanja / Scherer, Klaus R. 2005. The role of intonation in emotional expressions. *Speech Communication*, 46/3-4, 252–267.

Boersma, Paul / Weeninck, David 1996. Praat, a system for doing phonetics by computer. Technical Report 132, Institute of Phonetic Sciences of the University of Amsterdam, Amsterdam, Netherlands. <http://www.praat.org>.

Breitenstein, Caterina / Van Lancker, Diana / Daum, Irene 1986. The contribution of speech rate and pitch variation to the perception of vocal emotions in a German and an American sample. *Cognition & Emotion*, 15/1, 57–79.

Bulut, Murtaza / Narayanan, Shrikanth S. / Syrdal, Ann K. 2002. Expressive speech synthesis using a concatenative synthesizer. In *7th International Conference on Spoken Language Processing (ICSLP 2002)*, pages 1265–1268, Denver, CO, USA, September 2002.

Bulut, Murtaza / Busso, Carlos / Yildirim, Serdar / Kazemzadeh, Abe / Lee, Chul Min / Lee, Sungbok / Narayanan, Shrikanth S. 2005. Investigating the role of phoneme-level modifications in emotional speech resynthesis. In *9th European Conference on Speech Communication and Technology (Interspeech'2005 - Eurospeech)*, pages 801–804, Lisbon, Portugal, September 2005.

Bulut, Murtaza / Lee, Sungbok / Narayanan, Shrikanth S. 2007. Analysis of emotional speech prosody in terms of part of speech tags. In *Interspeech 2007 - Eurospeech*, pages 626–629, Antwerp, Belgium, August 2007.

Bulut, Murtaza / Narayanan, Shrikanth S. 2008. On the robustness of overall F_0-only modification effects to the perception of emotions

in speech. *Journal of the Acoustical Society of America*, 123/6, 4547-4558, June 2008.

Burkhardt, Felix / Paeschke, Astrid / Rolfes, Miriam / Sendlmeier, Walter F. / Weiss, Benjamin 2005. A database of German emotional speech. In *9th European Conference on Speech Communication and Technology (Interspeech'2005 - Eurospeech)*, pages 1517–1520, Lisbon, Portugal, September 2005.

Burkhardt, Felix / Sendlmeier, Walter F. 2000. Verification of acoustical correlates of emotional speech using formant-synthesis. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, pages 151–156, Newcastle, Northern Ireland, UK, September 2000.

Busso, Carlos / Narayanan, Shrikanth S. 2006. Interplay between linguistic and affective goals in facial expression during emotional utterances. In *7th International Seminar on Speech Production (ISSP 2006)*, pages 549–556, Ubatuba-SP, Brazil, December 2006.

Busso, Carlos / Lee, Sungbok / Narayanan, Shrikanth S. 2008. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *IEEE Transactions on Audio, Speech and Language Processing*, In press 2008.

Cahn, Janet E. 1990. The generation of affect in synthesized speech. *Journal of the American Voice I/O Society*, 8, 1–19, July 1990.

Clavel, Chloé / Vasilescu, Ioana / Devillers, Laurence / Richard, Gaël / Ehrette, Thibaut 2008. Fear-type emotion recognition for future audio-based surveillance systems. *Speech Communication*, 50/8, 487-503, June 2008.

Cover, Thomas M. / Thomas, Joy A. 2006. *Elements of Information Theory*. Wiley-Interscience, New York, NY, USA, 2006.

Cowie, Roddy / Douglas-Cowie, Ellen / Tsapatsoulis, Nicolas / Votsis, George / Kollias, Stefanos / Fellenz, Winfried / Taylor, John G. 2001. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18/1,32–80, January 2001.

Cowie, Roddy / Cornelius, Randolph R. 2003. Describing the emotional states that are expressed in speech. *Speech Communication*, 40/1-2, 5–32, April 2003.

Douglas-Cowie, Ellen / Campbell, Nick / Cowie, Roddy / Roach, Peter 2003. Emotional speech: Towards a new generation of databases. *Speech Communication*, 40/1-2,33–60, April 2003.

Duda, Richard O. / Hart, Peter E. / Stork, David G. 2000. *Pattern Classification*. Wiley-Interscience, New York, NY, USA, 2000.

Ekman, Paul / Friesen, Wallace V. / O'Sullivan, Maureen / Chan, Anthony / Diacoyanni-Tarlatzis, Irene / Heider, Karl / Krause, Rainer / LeCompte, William A. / Pitcairn, Tom / Ricci-Bitti, Pio E. / Scherer, Klaus R. / Tomita, Masatoshi / Tzavaras, Athanase 1987. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology*, 53:712–717, October 1987.

Ekman, Paul / Friesen, Wallace V. 1978. *Facial Action Coding System: A Technique for Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, CA, USA.

Grimm, Michael / Kroschel, Kristian / Mower, Emily / Narayanan, Shrikanth S. 2007. Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*, 49/10-11, 787–800, October-November 2007.

Izard, Carroll E. 1977. *Human Emotions*. Plenum Press, New York, NY, 1977.

Jurafsky, Daniel / Martin, James H. 2000. *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Prentice-Hall, Inc., Upper Saddle River, New Jersey, 2000.

Juslin, Patrik N. / Laukka, Petri 2003. Communication of emotions in vocal expression and music performance: different channels, same code? *Psychological Bulletin*, 129/5, 770–814, September 2003.

Klabbers, Esther / Van Santen, Jan P.H. 2004. Clustering of foot-based pitch contours in expressive speech. In *5th ISCA Speech Synthesis Workshop*, pages 73–78, Pittsburgh, PA, USA, June 2004.

Ladd, D. Robert / Silverman, Kim E. / Tolkmitt, Frank / Bergmann, Günther / Scherer, Klaus R. 1985. Evidence for the independent function of intonation contour type, voice quality, and F_0 range in signaling speaker affect. *Journal of the Acoustical Society of America*, 78/2, 435–444, August 1985.

Lee, Chul Min / Yildirim, Serdar / Bulut, Murtaza / Kazemzadeh, Abe / Busso, Carlos / Deng, Zhigang / Lee, Sungbok / Narayanan, Shrikanth S. 2004. Emotion recognition based on phoneme classes. In *8th International Conference on Spoken Language Processing (ICSLP 04)*, pages 889–892, Jeju Island, Korea, October 2004.

Lee, Sungbok / Yildirim, Serdar / Kazemzadeh, Abe / Narayanan, Shrikanth S. 2005. An articulatory study of emotional speech production. In *9th European Conference on Speech Communication and Technology (Interspeech'2005 - Eurospeech)*, pages 497–500, Lisbon, Portugal, September 2005.

Liberman, Mark / Davis, Kelly / Grossman, Murray / Martey, Nii / Bell, John 2002. Emotional prosody speech and transcripts, 2002. Linguistic Data Consortium.

Lieberman, Philip / Michaels, Sheldon B. 1962. Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech. *Journal of the Acoustical Society of America*, 34/7, 922–927, July 1962.

Montero, Javier M. / Gutierrez-Arriola, Juana M. / Colás, José / Enriquez, Emilia / Pardo, José Manuel 1999. Analysis and modelling of emotional speech in Spanish. In *International Congress of Phonetic Sciences (ICPhS 1999)*, pages 957–960, San Francisco, CA, USA, August 1999.

Moulines, Eric / Charpentier, Francis 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9/5-6, 453–467, December 1990.

Murray, Iain R. / Arnott, John L. 1993. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustical Society of America*, 93/2,1097–1108, February 1993.

Murray, Iain R. / Arnott, John L. 1995. Implementation and testing of a system for producing emotion-by-rule in synthetic speech. *Speech Communication*, 16/4, 369–390, June 1995.

Murray, Iain R. / Edgington, Mike D. / Campion, Diane / Lynn, Justin 2000. Rule-based emotion synthesis using concatenated speech. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, pages 173–177, Newcastle, Northern Ireland, UK, September 2000.

Paeschke, Astrid 2004. Global trend of fundamental frequency in emotional speech. In *Speech Prosody (SP 2004)*, pages 671–674, Nara, Japan, March 2004.

Paeschke, Astrid / Kienast, Miriam / Sendlmeier, Walter F. 1999. F_0-contours in emotional speech. In *Proceedings of the 14th International Conference of Phonetic Sciences (ICPh 1999)*, pages 929–932, San Francisco, CA, USA, August 1999.

Pantic, Maja / Rothkrantz, Leon J.M. 2003. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91/9, 1370–1390, September 2003.

Paul, Douglas B. / Baker, Janet M. 1992. The design for the Wall Street Journal-based CSR corpus. In *2th International Conference on Spoken Language Processing(ICSLP 1992)*, pages 899–902, Banff, Alberta, Canada, October 1992.

Picard, Rosalind W. 1997. *Affective Computing*. MIT Press, Cambridge, MA, USA, 1997.

Raux, Antoine / Black, Alan W. 2003. A unit selection approach to F_0 modeling and its application to emphasis. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2003)*, pages 700–705, St Thomas, US Virgin Islands, November-December 2003.

Roach, Peter 2000. Techniques for the phonetic description of emotional speech. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, pages 53–59, Newcastle, Northern Ireland, UK, September 2000.

Scherer, Klaus R./ Ladd, D. Robert / Silverman, Kim E. 1984. Vocal cues to speaker affect: Testing two models. *Journal of the Acoustical Society of America*, 76/5,1346–1356, November 1984.

Scherer, Klaus R. 1986. Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, 99/2, 143–165, March 1986.

Scherer, Klaus R. 2003. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40/1-2, 227–256, April 2003.

Schröder, Marc / Cowie, Roddy / Douglas-Cowie, Ellen / Westerdijk, Machiel / Gielen, Stan 2001. Acoustic correlates of emotion dimensions in view of speech synthesis. In *European Conference on Speech Communication and Technology (Eurospeech)*, volume 1, pages 87–90, Aalborg, Denmark, September 2001.

Schuller, Bjorn / Seppi, Dino / Batliner, Anton / Maier, Andreas / Steidl, Stefan 2007. Towards more reality in the recognition of emotional speech. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007)*, volume 4, pages 941–944, Honolulu, HI, USA, April 2007.

Traunmüller, Hartmut 2005. Speech considered as modulated voice. revised manuscript (last retrieved March 9, 2008), 2005.

Van Gogh, Vincent 1889. Letter 603 to Theo van Gogh written on 6 July 1889. Retrieved, July 7th, 2008. Web site: <http://www.vggallery.com/>.

Wang, Haibo / Li, Aijun / Fang, Qiang 2005. F_0 contour of prosodic word in happy speech of Mandarin. In Tao, J. / Tan, T. / Picard, R.W. (eds), *Affective Computing and Intelligent Interaction (ACII 2005), Lecture Notes in Artificial Intelligence 3784*, pages 433–440. Springer-Verlag Press, Berlin, Germany, November 2005.

Whissel, Cynthia M. 1989. The dictionary of affect in language. In R. Plutchik and H. Kellerman, editors, *Emotion: Theory, research and experience: Vol.4, The measurement of emotions*. Academic Press, New York, NY, USA, 1989.

Yildirim, Serdar / Bulut, Murtaza / Lee, Chul Min / Kazemzadeh, Abe / Busso, Carlos / Deng, Zhigang / Lee, Sungbok / Narayanan, Shrikanth S. 2004. An acoustic study of emotions expressed in speech. In *8th International Conference on Spoken Language Processing (ICSLP 04)*, pages 2193–2196, Jeju Island, Korea, October 2004.