The Roles of Pitch Contours in Differentiating Anger and Joy in Speech

Suthathip Chuenwattanapranithi, Yi Xu, Bundit Thipakorn, and Songrit Maneewongvatana

***Abstract*—**The present study is an attempt to use parameters that specify detailed pitch contours to improve the performance of classifying joy and anger emotions in speech. Three parameters, pitch range, speed of pitch change and peak alignment, were taken from accented and unaccented syllables in anger and joy speech samples. We also explore the possibility that alternative strategies of specifying these parameters may be used in the same emotion. The results show that, by using these syllable-based parameters and by allowing for alternative strategies, the performance of differentiating anger and joy was improved by 16%. Moreover, we showed that the new measurements could be used in predicting the pitch contours of accented syllables in anger and joy speech. This suggests the possibility of using these parameters to generate pitch contour related to emotion in speech synthesis.

***Keywords*—**emotional speech, multi-strategy classification, pitch contour, predicted line of pitch contour

I. Introduction

HUMAN speech conveys not only linguistic information but also information about the identity, age, geographical origin, attitude, and emotional state of the speaker. The nonlinguistic information is as important as linguistic information for human communication. In this work we concentrate on the emotional state embedded in human speech signal. Emotional speech recognition and synthesis play important roles in many applications. The direct uses include human machine interaction, entertainment, and business (call center). For indirect uses, the recognition of emotional features in speech can improve the performance of speech recognition systems [1].

Psychologists have been representing emotional state on the activation-evaluation space [2]. These values represent the degree of arousal and pleasantness, respectively. The experimental results from previous works suggest that some acoustic features are associated with general characteristics of emotional rather than a specific emotional state [1]. For example, high-activation emotions such as anger and joy have similar characteristics of greater loudness, higher pitch, and faster speed than low-activation emotions such as sadness. Many researchers suggest that grouping emotions based on high and low degrees of activation improve the performance of emotion recognition. Few works have concentrated on distinguishing emotions between the high- and low-evaluation emotions such as anger and joy.

In a preliminary experiment, we developed an emotional speech classifier using three parameters, mean pitch, intensity, and speech rate, which are widely used in other works. The performance of this classifier (Table I) was quite low for differentiating between anger and joy as compared to the evaluation by human (Table II). This suggests that these three parameters are not sufficient to represent the differences between anger and joy.

TABLE I
Performance of The Previous Classifier

Stimuli	Response (%)	
	Anger	Joy
Anger	69.23	30.77
Joy	58.33	41.67

The average accuracy rate is 55.45%

TABLE II
Performance of Human Evaluation

Stimuli	Response (%)		
	Anger	Joy	Sadness
Anger	78.48	20.25	1.27
Joy	25.33	70.67	4.00
Sadness	22.35	2.35	75.29

A possible reason for the low recognition rates in our preliminary experiment is that, because anger and joy are both high-activation emotions, their intensity and speech rate are likely very similar. It is possible, however, that intonation patterns are more useful than other parameters in differentiating anger and joy [3]. Some previous works have addressed the roles of intonation, especially which of pitch contour patterns in manifesting emotional speech [4], [5]. The intonation parameters that have been used before were rather simplistic, which may be partially responsible for the low classification rates so far. To find more sophisticated parameters, we considered the recently proposed PENTA model of tone and intonation [6]. The PENTA model assumes that the most basic tonal and intonation units are underlying pitch targets defined as simple linear functions, and the surface F0 is the result of speakers' sluggish articulatory approximation of these targets. These underlying targets are

Manuscript received January 25, 2006.

S. Chuenwattanapranithi is with the Department of Computer Engineering, King Mongkut's University of Technology Thonburi, Bangkok, Thailand and she is an affiliated research student at the Department of Phonetics and Linguistics, University College London, London, UK (corresponding author to provide phone: (+66)0-2470-9083; fax: (+66)0-2872-5050; e-mail: fay@phonetics.ucl.ac.uk, chuenwattana@yahoo.com).

Yi Xu is with the Department of Phonetics and Linguistics at University College London, London, UK and he works with Haskins Laboratories, New Harven, CT, USA (e-mail: yi@phonetics.ucl.ac.uk).

B. Thipakorn is with the Department of Computer Engineering at King Mongkut's University of Technology Thonburi, Bangkok, Thailand (e-mail: bundit@cpe.kmutt.ac.th).

S. Maneewongvatana is with the Department of Computer Engineering at King Mongkut's University of Technology Thonburi, Bangkok, Thailand (e-mail: songrit@cpe.kmutt.ac.th).

assumed to be associated with syllables rather than with words or phases. Following these assumptions, we obtain three parameters: pitch range, speed of pitch change, and F0 peak alignment from both accented and unaccented syllables in emotional speech samples. These parameters have been found to be useful in the study of lexical tones in Mandarin [7]. The PENTA model also assumes that tonal and intonational functions that are deliberately communicative, such as lexical tone, focus and interrogative meaning, are encoded as modifications of the target parameters as well as the speed of approximation of targets. At the same time, however, because emotions are not always deliberatively communicative, but often unintentionally revealed, emotional speech may not be consistently codified [6]. It is thus necessary to consider the possibility that the same emotion may be manifested in alternative ways in speech.

II. Speech Samples

The emotional speech samples in this experiment are the same set used in our preliminary experiment described in the first part (Table I). Sample speeches are obtained from English, German, French, Spanish, and Slovenian emotional speech databases which are publicly available [8], [9]. The emotional class of these emotional samples are already labeled and have been validated by human listeners. Each database consists of emotion words or sentences of male and female local speakers classified into three classes of emotion states: anger, joy, and sadness. We performed the experiment to validate the emotional class of each sample word or sentences. The criteria of choosing the sample words or sentences is that their meaning would not lead human subject to guess their emotional class. Therefore, the validation experiment was performed by 20 -Thai listeners who were not familiar with the languages of the databases. Each human subject was asked to identify the emotional state of each sample speech by hearing each word or sentence one at a time. There was a total of 60 sample words or sentences (20 samples for each emotional class). The total number of accented and unaccented syllables for anger and joy speech is 140 syllables (70 syllables for each emotional class). The performance of human evaluation of emotional speech is shown in Table II. The average accuracy rate is 74.8594%.

III. Measurements

An important step in building an emotion classifier is to select appropriate intonation units for analysis. According to the concept of intonation hierarchy [10], pitch contour of syllable is usually defined as the smallest component of the prosodic structure. The higher level pitch contours are composed of the smaller units such as pitch contour for words are the joint of syllable pitch contours. Analyzing pitch contour of each syllable has the advantage of probing more directly at the articulatory level of speech which presumably is affected by speakers' emotional states, while the complex intonation contours at the level of phrase or sentence are more directly related to the deliberate intonational functions. The measurements in this work were thus all taken at the level of the syllable rather than at the level of word or phrase. The accented syllables were extracted in order to measure the parameter values. According to the work of Tanja et al. [4], accented syllables are the ones which have local maximum pitch value located between two local minimum pitch values. Moreover, we also interested in the unaccented syllables, especially the final syllables. The following measurements were taken using Praat [11].

- Pitch range: Measured as the distance between maximum and minimum pitch values in one syllable (also known as excursion size). It is measured in semitone in order to make the data from individual speakers more comparable.

$$rF_0(semitone) = \frac{F_{0\max}}{F_{0\min}} \times \frac{1}{2^{1/12}} \quad (1)$$

Where rF_0 is F_0 range in semi tone unit $F_{0\max}$,and $F_{0\min}$ are the maximum and local minimum of F_0 in the syllable, respectively.

- Rising and falling strengths: First, velocity of F_0 curves were computed by taking the first order derivative of each F_0 curve (2). After that, linear regressions were performed to obtain linear equations with excursion size as the predictor (x) and velocity as the dependent variable (3). Then, the values of velocity (v) for two F_0 curves at a given excursion size (x) were compared. Fig. 1(a) shows the pitch contour of one accented syllable and the trajectory of its first derivative. Fig. 1(b) are the linear approximation of rising and falling strength.

$$v(t) = \frac{d}{dt} f_0(t) \quad (2)$$

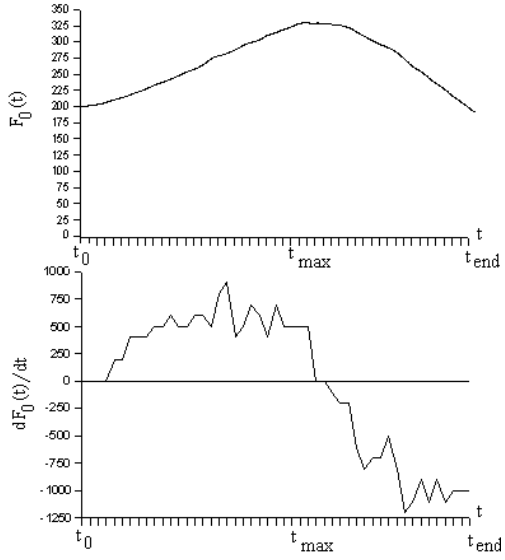
$$v(x) \approx ax + b \quad (3)$$

Where $v(t)$ is the velocity curve, $f_0(t)$ is the F_0 contour, and $v(x)$ is the linear approximated velocity when excursion size is equal to x.

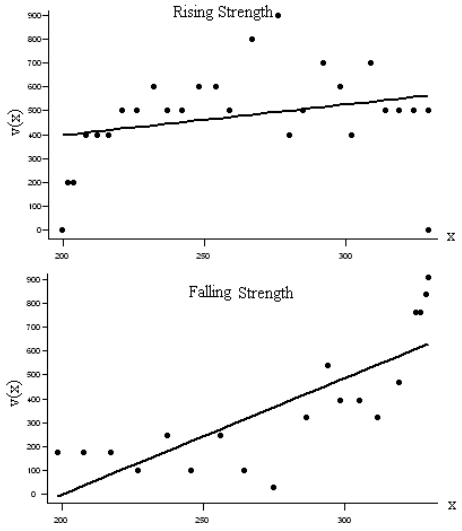
- Peak alignment: The proportion of time taken to reach maximum pitch value relative to syllable duration.

$$Pk = \frac{(t_{max} - t_0)}{(t_{end} - t_0)} \quad (4)$$

Where Pk is the peak alignment, t_{max} is the time of the maximum F_0, t_0 is the time of the first local minimum F_0, and t_{end} is the time of the second local minimum F_0.



(a)



(b)

Fig. 1 Pitch contour, velocity curve (a), and linear approximated velocity curve (b)

IV. Analysis and Results

After measuring the parameters from all samples, about 65% of them (56 pairs of anger and joy samples) were used as training data. As hypothesized in the Introduction, due to the non-deliberate nature of emotion in speech, the same emotional state may have alternative acoustic manifestations. For accented syllables, the K-mean algorithm involved the generating of possible strategies by using the diferent values of F_0 range and peak alingment between anger and joy as the features for clustering. Fig. 2 shows two clusters of the differences of feature values between anger and joy speech for accented syllables. For the unaccented syllables, the pitch target is used in clustering.

A differential classification method was therefore designed as shown in Fig. 3., in which features 1, 2, and 5 are the prominent ones. Paired t tests were performed to test if there are differences between pairs of anger and joy syllables uttered by the same speaker. The results from this test show that there are significant difference in the values of specific parameters in each strategy. The significance levels are shown in Table III. In the case of accented syllables, strategy 1 seems to be the most prominent. In this strategy, speakers utter with wider pitch range for joy than anger, and also higher strength for both F_0 rises and falls. For strategy 2 there are no difference in pitch range between the two emotions but strength and peak alignment are significantly different. Joy has an earlier peak alignment, higher rising, and lower falling strength than anger in the same syllable. The mean values of F_0 range and peak alingment used as the features for clustering for anger and joy in strategy 1 and 2 are shown in Table IV. Some speakers lower pitch near the end of the syllable for anger but not for joy. So we put them in strategy 3. For the case of unaccented syllables (mostly in final syllables), strategy 5 is dominant. Speakers produced low pitch for both anger and joy but the strength of falling pitch is higher in anger then in joy.

TABLE III
THE SIGNIFICANCE LEVEL (P-VALUE) FOR EACH PARAMETER

Strategy No.	Parameter			
	Pitch Range (Semi Tone)	Rising Strength	Falling Strength	Peak Alignment
1	0.0000 (Joy > ang)	0.0000 (Joy > ang)	0.0002 (Joy > ang)	0.2013
2	0.3098	0.0033 (Joy > ang)	0.0232 (Joy < ang)	0.0016 (Joy < ang)
5	0.6400	-	0.0374 (Joy < ang)	-

TABLE IV
MEAN VALUES OF PITCH RANGE AND PEAK ALIGNMENT FOR ACCENTED SYLLABLES

Strategy No.	Emotion	Mean of pitch range $\bar{x}_{pr}$	Mean of peak alignment $\bar{x}_{pa}$
1	Anger	1.2805	0.3091
	Joy	1.7605	0.3687
2	Anger	1.3953	0.4113
	Joy	1.5553	0.2813

Another 35% of speech samples (30 pairs of anger and joy samples) were used for testing, which was done with Bayesien classification. Moreover, discriminant scores were used in order to make decisions in cases where a sample was assigne to both emotions by different strategies. The experimental results are shown in Table V.The average accuracy rate is 71.795%.

TABLE V
PERFORMANCE OF THE PROPOSE CLASSIFIER

Stimuli	Response (%)	
	Anger	Joy
Anger	76.92	23.08
Joy	33.33	66.67

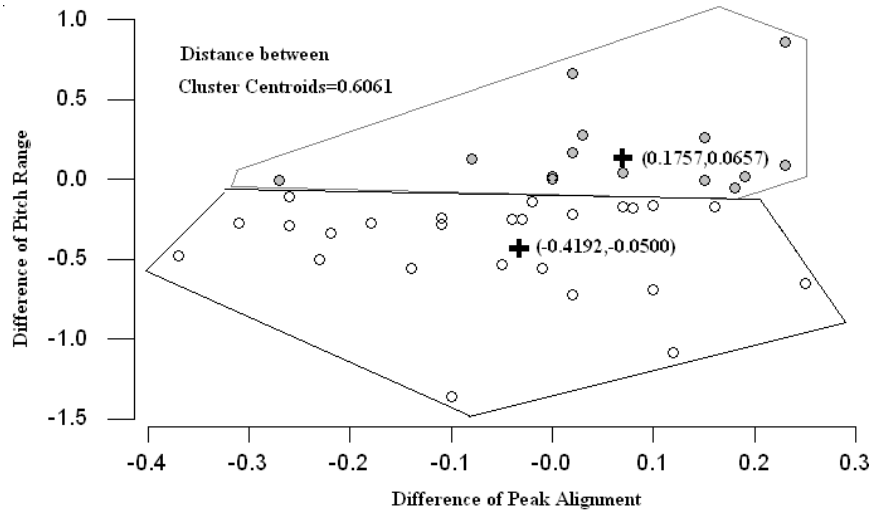


Fig. 2 Two Clusters for accented syllables

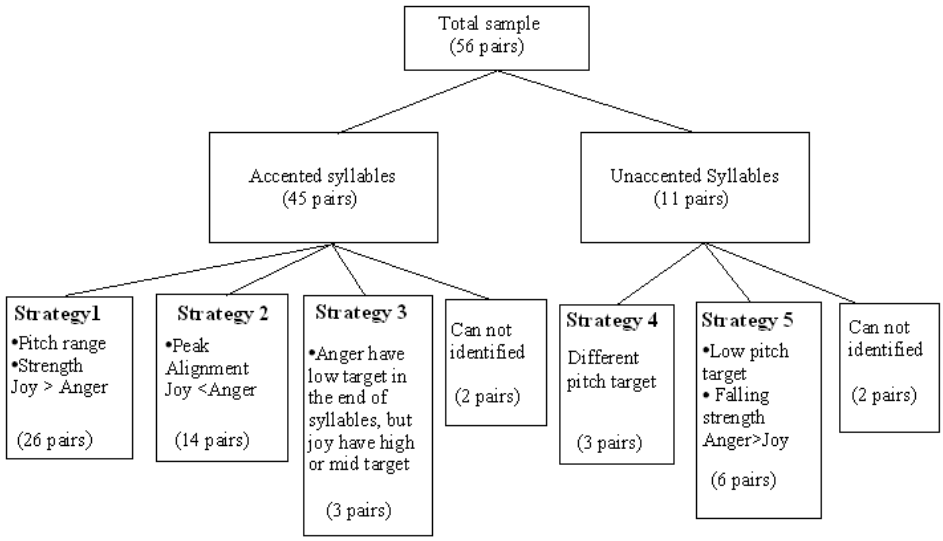


Fig. 3 Multi-strategy classification method for differentiating anger and joy in speech

V. Prediction of Pitch Contours

To further verify the efficacy of the specified pitch contour parameter, and the new classification method (multi-strategy classification), we used the mean values of the measurements used in strategies 1 and 2 (as presented in Table IV)to approximate the pitch contours of the accented syllables. The predicted line of pitch contour is calculated by linear equations
as in (5)-(7). The components of predicted line are presented in Fig. 4.

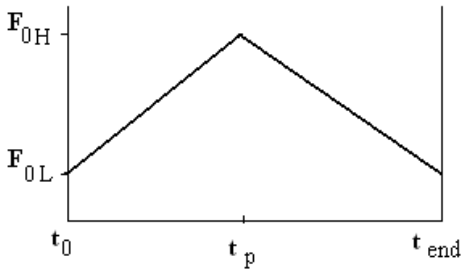


Fig. 4 Predicted line of pitch contour for accented syllables

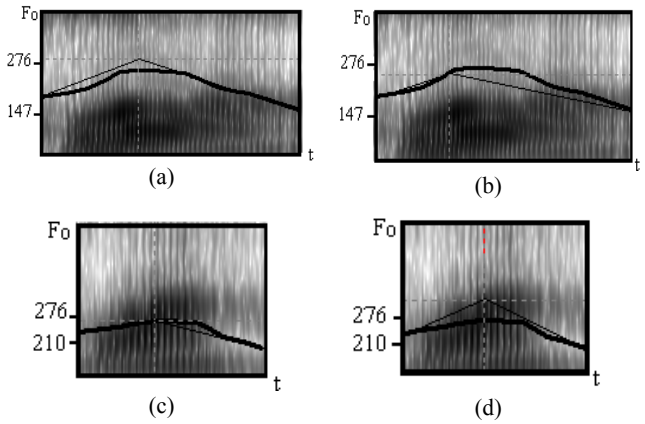
$$F_{0H} = F_{0L} \times 2^{1/12} \times \bar{x}_{pr} \quad (5)$$

$$t_p = \bar{x}_{pa} \times (t_{end} - t_0) + t_0 \quad (6)$$

$$p(t) = \begin{cases} \dfrac{F_{0H} - F_{0L}}{t_p - t_o} \times (t - t_0) + F_{0L}, & t_0 \le t \le t_p \\ -\dfrac{F_{0H} - F_{0L}}{t_{end} - t_p} \times (t - t_p) + F_{0H}, & t_p \le t \le t_{end} \end{cases} \quad (7)$$

Where F_{0H} and F_{0L} are the maximum and minimum of F_0 in the syllable, respectively, $\bar{x}_{pr}$ is the mean value of F_0 range for anger or joy emotion in each strategy, t_0, t_p and t_{end} are the beginning time, time of the peak F_0, and the ending time of the syllable, respectively, $\bar{x}_{pa}$ is the mean value of peak alignment for anger or joy emotion in each strategy, $p(t)$ is the predicted line. The values of $\bar{x}_{pr}$ and $\bar{x}_{pa}$ for anger and joy in strategy 1 and 2 are shown in Table IV.

Some of the results are illustrated in Fig. 5. The comparisons between predicted lines calculated with the new method (thin line) and actual pitch contour (thick line) for two accented syllables ((a)- (d), and (e)- (h)) are presented. For the first syllable, Fig. 5(a) and (b) are joy syllables predicted based on strategy 1 and 2, respectively. Fig. 5(c) and (d) are anger syllables predicted based on strategy 1 and 2, respectively. For the second syllable, Fig. 5(e) and (f) are joy



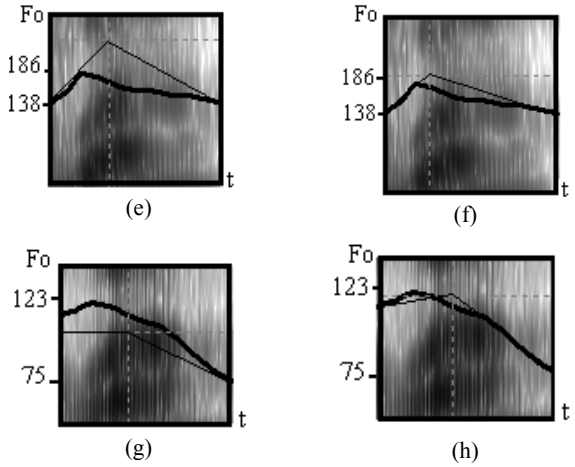


Fig. 5 Comparisons between predicted lines calculated with the new method and actual pitch contours

TABLE VI
ROOT MEANS SQUARE ERROR OF THE PREDICTED LINES OF FIG. 5

Syllable	Emotion	RMS Error Strategy 1 (Hz)	RMS Error Strategy 2 (Hz)
1	Anger	60.19 (Fig. 5(a))	116.14 (Fig. 5(b))
	Joy	36.91 (Fig. 5(c))	70.98 (Fig. 5(d))
2	Anger	480.15 (Fig. 5(e))	88.32 (Fig. 5(f))
	Joy	69.45 (Fig. 5(g))	31.05 (Fig. 5(h))

syllables predicted based on strategy 1 and 2, respectively. Fig. 5(g) and (h) are anger syllables predicted based on strategy 1 and 2, respectively. The root means square error of the predicted lines and actual pitch contours are shown in Table VI. The comparisons between predicted lines and actual pitch contours show that the actual pitch contours can be fitted with any one of two possible approximated lines generated from strategies 1 and 2. Fig. 5 (a)- (d) show that predicted lines calculated with strategy 1 are best for approximating pitch contours of both anger and joy. Fig. 5 (e)- (h) show that strategy 2 gives better results than strategy 1 for anger and joy. This result also confirms the results of the previous section that there are different manifestations of the same emotion, whether joy or anger. Moreover, the best-fit lines indicate good performance in predicting the actual pitch contours. It is therefore possible to use these measurements both for classifying emotions and for generating pitch contours related to emotional speech.

VI. DISCUSSION

The results of this experiment show that it is possible to classify high-activation emotions based on intonation contours. This can be done with a new method that treats the

same emotion as having variable manifestations, based on the assumption that emotions in speech are not deliberately communicative, and thus less codified as the more linguistic intonational functions. The different manifestations can be specified by various strategies. The first and the most dominant strategy we have found involves both pitch range and strength. With this strategy, pitch range is wider and strength is higher for joy than for anger. The second strategy involves F_0 peak alignment. With this strategy we have found that the maximum pitch is reached earlier for joy than for anger.

The results also show that pitch range, strength and peak alignment are not always directly related. As we can see from Table III, while both rising strength and lowering strength are significantly higher for joy than for anger, in strategy 5 anger has higher lowering strength than joy. This seems to suggest that there is an overall greater pitch raising trend in joy speech than in anger speech, and there is a greater pitch lowering tendency in anger speech than in joy speech. These trends are interpretable in terms of the theory of biological code proposed by Ohala [12]. From this theory, we may speculate that lowering pitch is consistent with sounding assertive and authoritative, which is likely more closely associated with anger than with joy. Table III also demonstrates the effectiveness of strategy 5, which shows that emotional information is conveyed not only by accented syllables but also by unaccented syllables, especially the final syllables.

The accuracy rates of classification show that the three measurements taken in the present study, namely, pitch range, speed of pitch change and peak alignment, which used with the muti-strategy method, are more effective in differentiating anger and joy than previously used parameters such as pitch, intensity, and speech rate. The performance was improved by about 16%. Moreover, the greater specificity of the new measurements gives them a better prospective in being used in simulating emotion in speech synthesis.

VII. Conclusion

This paper focuses on the roles of intonation, especially pitch contours to classify anger and joy in speech. Three new measurements, pitch range, speed of pitch change, and peak alignment, were obtained from anger and joy samples at the syllable level. Based on the non-deliberate nature of emotions, we allowed the classification process to identify multiple strategies associated with each emotion. The two dominant strategies related to the accented syllables effectively separated joy from anger: (a) making wider pitch range and doing it at greater strength, and (b) reaching the F_0 peaks earlier. The dominant strategy for unaccented syllables that separated anger from joy is having greater pitch falling strength. The much improved classification rate indicates that the new method can improve the performance of emotion classifiers for speech. Moreover, the F_0 contours predicted with the new measurements have a good match to the actual F_0 contours in accented syllables. This suggests applicability of the new measurements in generating of emotional speech in synthesis.

Acknowledgment

We would like to thank Dr. Tenja Banziger, Department of Psychology, University of Geneva for the permission of using German emotional speech samples.

References

[1] T.L. Nwe, S.W. Foo, and L.C. De Silva, Speech emotion recognition using hidden Markov models, Speech Communication, Vol. 41, Issue 4, pp. 603-623, 2003.

[2] C.M. Whissel, "The dictionary of affect in language", In: R. Plutchik and H. Kellerman (Eds.)Emotion: Theory, Research and Experience: Vol. 4,The Measurement of Emotions, Academic Press New York, pp.113-131, 1989.

[3] X. Sun, The deternination, analysis, and synthesis of fundamental frequency, PhD. Desertation, 2002.

[4] T. Bänziger and K.R. Scherer,The role of intonation in emotional expressions,Speech Communication, Vol. 46, Issues 3-4, pp.252-267, 2005.

[5] K. Hirose, K. Sato, Y. Asano, and N. Minematsu, Synthesis of F0 contours using generation process model parameters predicted from unlabeled corpora: application to emotional speech synthesis ,Speech Communication, Volume 46, Issues 3-4,pp. 385-404, 2005.

[6] Y. Xu, Transmitting tone and intonation simultaneously -The parallel encoding and target approximation (PENTA) model, Proceedings of International symposium on tonal aspects of language,pp. 215-220, 2004.

[7] Y. Xu, Effects of tone and focus on the formation and alignment of F0 contours. Journal of Phonetics, Vol. 27, pp. 55-105.1999.

[8] R.Banse, K.R.Scherer, Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology 70, pp. 614–636, 1996.

[9] Emotional Speech Group, DSPLAB, University of Maribor web page: http://wwwbox.uni-mb.si/eSpeech/

[10] A. Botinis., B. Granström ,and B. Möbius, Developments and paradigms in intonation research. Speech Commun. 33, pp. 263-296, 2001.

[11] P. Boersma, D.J.M. Weenink, Praat, a System for Doing Phonetics by Computer, Version 3.4 (132). Institute of Phonetic Sciences of the University of Amsterdam, Amsterdam, 1996.

[12] J. J. Ohala, An ethological perspective on common cross-language utilization of F0 of voice. Phonetica 41, pp. 1-16, 1984.