RECOGNITION FOR SYNTHESIS: AUTOMATIC PARAMETER SELECTION FOR RESYNTHESIS OF EMOTIONAL SPEECH FROM NEUTRAL SPEECH

Murtaza Bulut, Sungbok Lee and Shrikanth Narayanan**

University of Southern California, Los Angeles, CA
Department of Electrical Engineering, *Also, Department of Linguistics

ABSTRACT

One of the biggest challenges in emotional speech resynthesis is the selection of modification parameters that will make humans perceive a targeted emotion. The best selection method is by using human raters. However, for large evaluation sets this process can be very costly. In this paper, we describe a recognition for synthesis (RFS) system to automatically select a set of possible parameter values that can be used to resynthesize emotional speech. The system, developed with supervised training, consists of synthesis (TD-PSOLA), recognition (neural network) and parameter selection modules. The experimental results show evidence that the parameter sets selected by the RFS system can be successfully used to resynthesize the input neutral speech as angry speech, demonstrating that the RFS system can assist in the human evaluation of emotional speech.

Index Terms— emotion resynthesis, automatic evaluation, neural network, recognition for synthesis

1. INTRODUCTION

Speech synthesis, and specifically emotional speech synthesis, is a challenging research topic. Two of the main challenges are that (1) there are numerous parameter values that can be selected during the generation of pitch, duration, and energy contours, and that (2) human evaluators are needed to evaluate synthesizers' performances. The need for human subjects requires that evaluation experiments be carefully designed to minimize the cost, both in terms of time and resources. At the same time, in order to find the best balance between different parameter values, many combinations need to be tested. Clearly, there is a trade off between the design requirements and the cost.

In this paper we address the synthesis of angry and happy emotional speech and propose using an automatic emotion recognizer as a preprocessing step to narrow down the size of the evaluation set before it is presented to human raters. The proposed system, trained from labeled emotional data, consists of a prosody modification module which generates a large number of synthetic utterances that are then evaluated using a neural network (NN) emotion recognizer. The output of the recognizer is used to select the parameter combinations performing consistently well, and only these modifications are submitted for evaluations with human subjects.

The emotion characteristics of speech can be partly associated with the changes in the prosody (pitch, duration, and energy) parameters [1, 2] and partly with the spectral characteristics of speech [3, 4]. However, as explained in [2, 5] it should be noted that human emotion perception is a complex process which involves many other factors. In this paper the concentration is just on the prosody parameters.

For synthesis (which in this paper we use as a synonym to resynthesis) of some emotions, such as sadness, simple prosody modification rules can be utilized. For instance, by lowering the F0 mean (by $\approx 30\%$), decreasing the F0 range (by $\approx 100\%$), and by increasing the duration (by $\approx 30\%$) it may be possible to synthesize a low activation and dull speech which would be perceived as sad (or bored, depressed, discontented, fed up, not in high spirit, or weak) under appropriate conditions. Conversely, if the F0 mean is increased beyond a certain level (more than 50% of its original value), speech which would also be perceived as sad (a different type of sadness, however) might be synthesized. This type of sadness can be described as an extreme sadness which has a distinct cry-like speech quality [3, 6].

Synthesis of high activation emotions, such as happiness and anger, is more challenging. Although it has been suggested that large F0 range and mean variations can be beneficial for the synthesis of these emotions [1], in practice it is not usually the case. In many cases, the large F0 mean and range modifications would add a cry-like (due to high pitch) and trembling (due to high jitter) quality to speech, which in turn would favor the perception of sadness. Clearly, there is a fine balance between different prosody modifications that are needed to successfully synthesize speech that will be perceived as angry or happy. In order to achieve this balance, many parameter value combinations need to be evaluated. Since using human raters for this process is costly, a more efficient technique is needed. For instance, a technique that will perform the evaluations automatically. Having such an automatic emotion evaluation system will be beneficial to find the best modification combinations specific to each sentence, to each emotion, and to each speaker.

In this paper we test how successfully machine recognition of emotions can be used to assist human recognition of emotions. Results for two emotions – happiness and anger – are reported. The goal of this work is to inform designs of Recognition for Synthesis (RFS) systems for the automatic evaluation of synthetic emotional speech.

2. RECOGNITION FOR SYNTHESIS (RFS) SYSTEM DESCRIPTION

The proposed algorithm consists of five main stages, which are briefly outlined in this section (see Fig. 1) and explained in more detail in the following sections. In stage one (see Sec. 2.1), pitch, duration, and energy of natural utterances are modified using an empirically selected set of parameter values. In stage two (Sec. 2.2), using a neural network emotion recognizer, these resynthesized utterances are classified into one of the angry, happy, sad or neutral emotion categories. In stage three (Sec. 3), the classification results are used to select the best parameters, which produce successful recognition results from a machine recognition perspective. In the fourth stage, using the selected parameters, input utterance prosody is modified,

1-4244-1484-9/08/$25.00 ©2008 IEEE

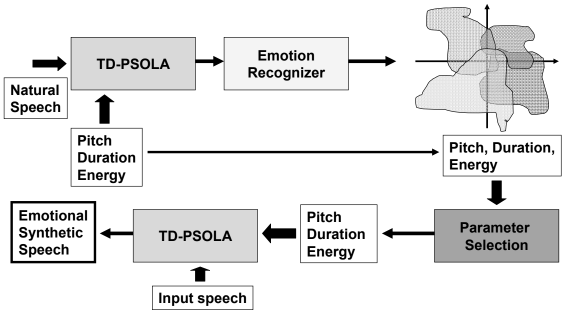

Fig. 1. Recognition for Synthesis (RFS) system architecture. The emotion recognizer is used to assess the emotional quality of the resynthesized utterances. Based on the results, parameters for synthesis are selected and applied to add emotional quality to the input speech.

and in the final stage (Sec. 4), listening experiments, assessing emotional quality, with human raters are conducted to determine the set of parameters that perform the best from a human recognition perspective.

2.1. Prosody modifications

For modification of the input speech, only prosody modifications were performed using the TD-PSOLA algorithm as implemented in the Praat software. The modifications were performed on the voiced (V) and unvoiced regions (U) of utterances by scaling the original utterance values by the factors listed below. The voiced and unvoiced region boundaries were automatically detected using the Praat software.

The tested scaling factors for duration modifications were [0.7, 1, 1.4], for energy modifications were [0.5, 1, 2], for F0 median modifications were [0.8, 1, 1.25], and for F0 range modifications were [0.4, 1, 2.5].

All of the possible modification factor combinations were tested. F0 modifications were applied only on voiced regions, while duration and energy modifications were applied on both voiced and unvoiced regions, independently.

The modified values were calculated by the multiplication of the modification factor and the original value. For example, if the F0 range, voiced region duration and unvoiced region duration modification factors were set to 2.5, 0.7, and 1.4 respectively, and the rest to 1, then for a given input utterance the F0 range would be increased by a factor of 2.5 (i.e., 150% of original). The voiced regions' durations would be decreased by factor of 0.7 (i.e., 30%) and the unvoiced regions' durations would be increased by 40%.

The range of the tested modification factors was chosen large to allow for broader coverage. In order to minimize the test time, the number of the tested factors was kept minimal. Note, however, that finer resolution of modification factors may be necessary, in the future, for more advanced and detailed analyses, and [probably, but not necessarily] for better selection of different parameter combinations. Increasing the number of the factors increases the number of the synthesized utterances exponentially. Even in this case, for example, for a given input utterance, 729 ($= 3^6$) new utterances were synthesized.

The most accurate and reliable way to evaluate the emotional content of synthesized utterances is through listening tests with many human raters. However this is not feasible due to the large size of the evaluating set. The emotion recognizer, described next, is proposed as a preprocessing step – to aid human listeners – for generating a more manageable evaluation set.

2.2. Automatic emotion recognition using neural networks

Neural networks (NN) are popular in machine learning applications because they are able to learn and model non-linear data with high success rates.

Our goal in this paper was to design an emotion recognition system capable of distinguishing four emotion (angry, happy, sad, neutral) types. For that purpose, we built a 1 hidden layer feed-forward neural network, with 31 inputs, 5 hidden units and 2 output units using Matlab's Neural Networks toolbox.

2.2.1. Input variables

The input variables (31 in total) used to train the NN included a variety of prosody parameters (calculated in Praat) as detailed below.

For the whole speech file the following parameters were calculated: (1) F0 mean, (2) F0 median, (3) F0 range, (4) F0 std, (5) F0 minimum, (6) F0 maximum, (7) Energy, (8) Intensity, (9) Duration, (10) 25% quantile of F0, (11) 75% quantile of F0.

Next, the voiced regions were extracted and concatenated together to generate a voiced-regions-only file. For this file, the following parameters were calculated: (12) F0 mean, (13) F0 median, (14) F0 range, (15) F0 std, (16) F0 minimum, (17) F0 maximum, (18) Energy, (19) Intensity, (20) Duration, (21) 25% quantile of F0, (22) 75% quantile of F0, (23) Intensity minimum, (24) Intensity maximum, (25) Relative intensity minimum position (= Intensity minimum time / Duration), (26) Relative intensity maximum position (= Intensity maximum time / Duration), (27) Intensity contour mean, (28) Intensity contour std.

Finally, the unvoiced regions were extracted and concatenated together to generate an unvoiced-regions-only file. For this file, (29) Intensity, (30) Duration, (31) Energy parameters were calculated.

Before training the system, all of the input variables were normalized to be in the $[0\ 1]$ range.

2.2.2. Output variables

Two output variables were used to represent each emotion. They were (1,1) for happy, (1,-1) for sad, (-1,-1) for angry, and (-1,1) for neutral. We used 2 dimensional output vectors because they provide a nice visualization of the four emotional spaces. In this case, each quadrant of the Cartesian coordinate system can be regarded as a distinct emotional space.

2.2.3. Neural network system design

For the construction of the NN system we used the Matlab's Neural Network toolbox. The neural network was trained with backpropagation using the following options: *trainrp*, *learngd*, *mse*, and 0.1 learning rate.

We experimented with a large number of NNs, by varying the number of hidden units, the number of hidden layers and type of the activation functions before deciding to use 1-hidden layer (with 5 units) with logarithmic sigmoidal (*logsig*) function and linear activation (*purelin*) at the output. This network was chosen because of its high performance, robustness and simple structure. The error function that was used for comparing different neural networks was the misclassification rate of the test emotional utterances.

2.2.4. *Training data and system performance*

The emotional data used for training the network is described in [7]. The training data consist of 408 utterances (102 for each emotion) and test data consist of 112 utterances (28 for each emotion) recorded by a professional actress in angry, happy, sad, or neutral emotions. The training and test sets were randomly split, and did not have any common utterances or sentences.

The NN network was trained and tested with 5 different training and test sets, and the recognition accuracies for these test sets were 81.25%, 74.14%, 70.00%, 82.00%, and 80.17%, averaging 77.77%. The average recognition accuracy for the 5 training sets was 94.43%. The confusion matrix summing the test results for these different runs is given in Table 1.

Emotion	Happy-NN	Sad-NN	Angry-NN	Neutral-NN
Happy	90 (62.93%)	17	3	33
Sad	7	109 (76.22%)	6	21
Angry	2	1	139 (97.20%)	1
Neutral	17	12	7	107 (74.83%)

Table 1. Confusion matrix of NN recognition results summing the test results of 5 different runs (112 test utterances in each run). Displayed are the number of the files, and percentages in parenthesis. Emotion-NN indicates the emotions recognized by the NN.

3. MODIFICATION FACTOR SELECTION

For further analysis we concentrated on test set 1, the NN recognition performance for which was 81.25%. First, the utterances falling inside the unit circle centered on the $[-1\ 1]$ point (which corresponds to the neutral output) were selected for further processing. These selected neutral utterances will be referred to as *SelNeu*. This set consisted of 21 utterances (out of possible 28).

To all of the utterances in *SelNeu*, the prosody modifications explained in Sec. 2.1 were applied. As a result 15309 (= 21 x 729) utterances were resynthesized. They will be referred to as *ModNeu*. (The NN performance on the *ModNeu* set was 14.64% happy, 22.59% sad, 7.28% angry, and 55.49% neutral, showing that the most of the modifications did not alter the neutral input emotion.)

Next, 5 neutral utterances[1] were randomly selected from the *SelNeu*. Let us call this set *EvalNeu*. For *EvalNeu* utterances, the modification factor combinations (*SelMod*) that made them classified as happy or angry were automatically determined. Note that naturally different modifications were selected for different target emotions and for different utterances.

In order to select among the large number of successful modifications, the *SelMod* modifications were sorted based on their performance on the *SelNeu* set. The procedure was as follows. First, the effect of each of the selected modifications (*SelMod*) on the utterances of *SelNeu* was determined. Next, the *SelMod* modifications were sorted in descending order based on the number of instances for which they produced the target emotion. A separate sorting was performed for each target emotion. After the sorting, the most consistently performing modifications were on the top of the stack.

The first five of the sorted *SelMod* modifications were selected to be used in the human listening experiments. These modifications will be represented as *h1, h2, h3, h4, h5* (for happy), and *a1, a2, a3, a4, a5* (for angry), and referred to as *BestSelMod* forth in the paper.

4. LISTENING TESTS

The utterances in set *EvalNeu* were modified according to the *BestSelMod* modification factors (different modification factors for every utterance and for every emotion) that were found in the previous step. As a result of these modifications 50 utterances (= {5 neutral utterances} x {5 *BestSelMod* modifications} x {happy, angry}) were resynthesized and presented to human raters for the listening test.

4.1. Listening test structure

A web based interface (a web page[2] prepared using Perl CGI) showing a table with 50 rows and 3 columns was used for listening tests. Three speech files were shown on each row. The first file was defined as a reference file and it was indicated that this utterance had *neutral* emotion with *confidence 5* (= the highest confidence). The other two files were two randomly selected modified versions of the reference file. One of these files was synthesized using the parameters selected (as explained in the previous section) for happy emotion, and the other for angry emotion. In a similar manner, all of the remaining utterances were presented on the same web page. The order of the utterances in each row and each column was randomly determined and it was different for every evaluator and every sentence.

Listeners were given 5 emotion options but were allowed to select only one of them. These options were *Neutral*, *Angry*, *Happy*, *Sad*, and *Other*. Note that although only the synthesis of anger and happiness was tested, the raters were presented with 5 selection options in order to be consistent with the evaluation of the natural speech of the same speaker done in [7]. Confidence level (for the emotion choice that the rater selects) was measured on a 5 point scale, 5 showing high confidence and 1 low confidence.

A total of 27 naive raters (10 female, 17 male) participated in the test. They were not given any detailed information about the nature of the test, except the fact that they needed to listen to some utterances and then select the emotions they perceived. All of the subjects had advanced English language skills and they were mostly engineering graduate students. Headphones were used by 19 of them, while the remaining 8 preferred loud speakers. The average test duration was approximately 10-15 minutes.

4.2. Listening test results

The test results are presented in Table 2, and Table 3 (matrices (1) and (2)). For each of the input utterances (*n1, n2, n3, n4, n5*) the most successful happy (h) (matrix (1)) or angry (a) (matrix (1)) modifications were determined based on the human raters' responses. The recognition percentages and confidence scores are shown in Table 2. The parameter factor values for these modifications are shown in Table 3. The Table 2 also shows the average recognition for the best 2, and the best 3 of happy (*h1, h2, h3, h4, h5*), and of angry (*a1, a2, a3, a4, a5*) modifications.

5. DISCUSSION

The results show that the proposed system can successfully select the modification parameters for *angry* emotion synthesis. For example (see Table 2) for *n5* one of the selected modifications (*a4*) by the system was confidently (4.0) perceived as angry by 92.31% of the listeners. Similarly, for $n1$, $n2$, $n3$, at least one of the modifications automatically selected by the system made the synthesized utterances perceived as *angry* above the chance rate (20%). The average values measured for the best 2 modifications (55.83%) and the best 3 modifications (46.35%) show that some of the other selected

[1] The utterances that were tested were the following: n1: *I am going shopping.*, n2: *Lucy ate all the chocolate.*, n3: *Mickey ate all the raisins.*, n4: *The cat's meow always makes my finger twitch.*, n5: *The saw is broken so chop the wood instead.*.

[2] http://sail.usc.edu/∼mbulut/cgi-bin/evalJul25/comp_evalNN.cgi

Sent.	Mod.	Neutral	Angry	Happy	Sad	Other
n1	h2	11.11 (4.3)	55.56 (3.9)	**18.52 (4.0)**	00.00 (–)	14.81 (3.3)
	a5	25.93 (4.1)	**62.96 (3.6)**	11.11 (3.7)	00.00 (–)	00.00 (–)
n2	h5	38.46 (3.6)	46.15 (3.5)	**3.85 (4.0)**	7.69 (3.5)	3.85 (4.0)
	a1	11.11 (4.7)	**85.19 (3.7)**	3.70 (5.0)	00.00 (–)	00.00 (–)
n3	h4	30.77 (4.1)	23.08 (3.3)	**15.38 (3.0)**	19.23 (3.2)	11.54 (3.7)
	a4	37.04 (3.8)	**51.85 (3.5)**	3.70 (4.0)	3.70 (4.0)	3.70 (3.0)
n4	h2	18.52 (4.0)	22.22 (3.0)	**33.33 (3.4)**	00.00 (–)	25.93 (4.2)
	a4	37.04 (4.4)	**22.22 (2.8)**	7.41 (2.0)	3.70 (4.0)	29.63 (3.4)
n5	h1	42.31 (3.9)	00.00 (–)	**19.23 (3.0)**	30.77 (3.1)	7.69 (3.0)
	a4	3.85 (3.0)	**92.31 (4.0)**	00.00 (–)	00.00 (–)	3.85 (3.0)
all	2 best-h	30.93 (3.9)	28.03 (–)	**16.92 (3.23)**	11.37 (–)	12.75 (3.5)
	2 best-a	27.29 (4.03)	**55.83 (3.41)**	6.00 (–)	4.16 (–)	6.72 (–)
all	3 best-h	32.29 (3.9)	21.92 (–)	**15.59 (3.3)**	18.64 (–)	11.56 (3.2)
	3 best-a	29.13 (4.0)	**46.35 (–)**	5.00 (–)	14.02 (–)	5.50 (–)

Table 2. Results of listening tests with humans. Recognition percentages (average confidence) are shown. The symbols *n1, n2, n3, n4, n5* represent neutral utterances that were modified. The symbols *(h2, h5, h4, h2, h1)*, and *(a5, a1, a4, a4, a4)* represent the best performing happy, and angry modifications, respectively.

$$\begin{array}{rc} & \begin{array}{cccccc} Fm & Fr & Vd & Ve & Ud & Ue \end{array} \\ \begin{array}{r} n1-h2 \\ n2-h5 \\ n3-h4 \\ n4-h2 \\ n5-h1 \end{array} & \begin{pmatrix} 1.0 & 2.5 & 0.7 & 0.5 & 0.7 & 2.0 \\ 1.0 & 2.5 & 0.7 & 0.5 & 1.4 & 1.0 \\ 1.0 & 2.5 & 1.4 & 2.0 & 0.7 & 1.0 \\ 1.0 & 2.5 & 1.0 & 0.5 & 0.7 & 0.5 \\ 1.0 & 2.5 & 1.4 & 0.5 & 0.7 & 1.0 \end{pmatrix} \end{array} \tag{1}$$

$$\begin{array}{rc} & \begin{array}{cccccc} Fm & Fr & Vd & Ve & Ud & Ue \end{array} \\ \begin{array}{r} n1-a5 \\ n2-a1 \\ n3-a4 \\ n4-a4 \\ n5-a4 \end{array} & \begin{pmatrix} 1.0 & 1.0 & 0.7 & 2.0 & 1.4 & 0.5 \\ 1.0 & 2.5 & 0.7 & 2.0 & 1.4 & 0.5 \\ 1.0 & 1.0 & 0.7 & 2.0 & 1.4 & 0.5 \\ 1.0 & 1.0 & 0.7 & 2.0 & 1.4 & 0.5 \\ 1.0 & 1.0 & 0.7 & 2.0 & 1.4 & 0.5 \end{pmatrix} \end{array} \tag{2}$$

Table 3. The modification factor values that worked the best. (Fm = F0 mean, Fr = F0 range, Vd = Voiced duration, Ve = Voiced energy, Ud = Unvoiced duration, Ue = Unvoiced energy).

modifications were also successful in converting neutral speech into angry speech.

As seen from the matrix 2 (in Table 3), for angry speech generation one need to decrease the voiced speech duration (Vd), and unvoiced speech energy (Ue), and increase voiced speech energy (Ve) and unvoiced speech duration (Ud).

For happy speech synthesis, only the result for the modification (*h2*) selected for neutral utterance $n4$ was above the chance level. In general, we observe that automatically selected modification for happy speech synthesis, caused the synthesized utterances to be perceived with wide range of emotions, mostly as neutral (e.g., $n2$, $n3$, $n5$), angry (e.g., $n1$, $n2$, $n3$, $n4$), or sad (e.g., $n3$, $n5$).

The low performance achieved for happy emotion can be attributed to several causes. First, examining the NN recognition results in Table 1 we note that natural happy utterances were recognized with 62.94% accuracy (cf. anger 97.20%), showing that the performance of the NN recognizer was moderate for happy emotion recognition. Later when selecting the best performing modifications, this might have caused the NN recognizer to misclassify some modifications which might be perceptually important for happiness, causing the *BestSelMod* for happy synthesis to be insufficient. Second, as examined in detail in [7], for the natural emotional speech, the speaker's expression of happiness was sometimes confusable with anger. A similar confusion between anger and happiness is observed in the results in Table 2, which indicates that an improved emotional database may perform better. Third, for synthesis of happiness simple modifications on voiced and unvoiced speech regions might not be sufficient. It can be expected that finer modifications taking the word, phrase and stress pattern structures into account would improve the results [4, 1, 2]. In addition F0 contour shape modifications can be also helpful. Also importantly, note that there are other factors beyond prosody that can influence emotion perception, e.g., spectral envelope characteristics [4, 3] and the transmission medium [2, 5]. Using just the prosody factors may be one of the causes of the limited performance observed for happy emotion. Note however that the idea (of using a recognizer to select data parts) itself is general.

The results show that there are clear differences between the automated emotion classification and human perception. For instance, many of the parameters selected by the system (even for angry speech) were not useful for the synthesis of the targeted emotion. Also, although not examined here, it may be the case that some perceptually important parameter combinations were not selected.

In order to better understand and model the relation between machine and human perception of emotions, in the future, a new RFS system comprising the modification of both prosody (duration, energy, F0 mean, range, and shape) and spectral parameters will be tested with more data. Also more comprehensive human listening tests will be conducted.

The end goal of the proposed technique is to use it as a feedback system in the emotional speech synthesizers to select and adjust the appropriate modification parameters.

6. CONCLUSION

Considering the wide range of possible modifications that can be applied on a speech signal to synthesize emotional speech, there is a need for a system that can select the parameters that are expected to perform well, thus narrowing down the sample set that needs to be evaluated by human raters. In this paper, such a system (recognition for synthesis (RFS)), combining emotion recognition and synthesis, is described.

The results show that the proposed RFS system is promising for selecting parameters for emotional speech resynthesis. Considering the significantly different performances for different emotions, and the differences observed between human and machine perception of emotions, however, at this stage we prefer to view the proposed automated evaluation more as a preprocessing step than a replacement to human evaluations. Our future research will be directed towards the design of more robust systems, more sophisticated parameter modifications, and experimenting with different parameter selection techniques and additional emotions.

7. REFERENCES

[1] R. Cowie, E. D. Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18(1), pp. 32–80, Jan. 2001.

[2] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40(1-2), pp. 227–256, 2003.

[3] M. Bulut, C. Busso, S. Yildirim, A. Kazemzadeh, C. M. Lee, S. Lee, and S. Narayanan, "Investigating the role of phoneme-level modifications in emotional speech resynthesis," in *Proc. of Eurospeech, Interspeech*, Lisbon, Portugal, 2005.

[4] M. Bulut, S. Narayanan, and A. K. Syrdal, "Expressive speech synthesis using a concatenative synthesizer," in *ICSLP*, Denver, CO, 2002.

[5] H. Traunmuller, "Speech considered as modulated voice," revised manuscript, 2005.

[6] F. Burkhardt and W. F. Sendlmeier, "Verification of acoustical correlates of emotional speech using formant-synthesis," in *ISCA Workshop on Speech and Emotion*, Northern Ireland, 2000.

[7] S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "An acoustic study of emotions expressed in speech," in *Proc. of ICSLP*, Jeju, Korea, Oct. 2004.