Real-time Emotion Detection System using Speech: Multi-modal Fusion of Different Timescale Features

Samuel Kim, Panayiotis G. Georgiou, Sungbok Lee, and Shrikanth Narayanan
Speech Analysis and Interpretation Lab. (SAIL)
University of Southern California, Los Angeles, USA.
kimsamue@usc.edu

***Abstract*— The goal of this work is to build a real-time emotion detection system which utilizes multi-modal fusion of different timescale features of speech. Conventional spectral and prosody features are used for intra-frame and supra-frame features respectively, and a new information fusion algorithm which takes care of the characteristics of each machine learning algorithm is introduced. In this framework, the proposed system can be associated with additional features, such as lexical or discourse information, in later steps. To verify the real-time system performance, binary decision tasks on angry and neutral emotion are performed using concatenated speech signal simulating real-time conditions.**

I. Introduction

The purpose of emotion studies in association with human computer interaction is to build a machine that better serves users' needs in a more natural and effective way; thus, emotion studies have gained great interest from the engineering community. In human-computer or human-human dialog systems, emotion recognition systems could provide users with improved services by being adaptive to their emotion. In many applications, furthermore, there is a growing need of real-time emotion detection systems [1].

Various studies in emotion recognition from speech have shown the benefits of different features and learning algorithms. Much of the studies, however, have focused on off-line system using acoustic information at the phone or utterance level. In contrast to off-line emotion recognition research, a real time system is constrained by lack of clear utterance segmentation, lexical information, speaker specific adaptation, and computational requirements. In this work, we propose a real time emotion classification system employing only a streaming speech source with no additional information and with speed consideration requirements. In addition, in the proposed system architecture we allow for additional information (as they become available) to be considered.

In the proposed real-time emotion detection system, we extract and fuse emotional information encoded at different timescales especially supra- and intra- frame level. The rationale behind this approach is that emotion is encoded at different levels of speech, such as supra-frame, intra-frame, and lexical level, and each timescale feature is complementary [2][3]. There have been many studies trying to classify emotion states model with each level of feature. Considering intra-frame features, Kwon *et al.* performed emotion recognition tasks using MFCC with various learning algorithms [4]. They argued that the intra-frame feature contains emotional information, even though it is not the best feature. Nwe *et al.* performed similar experiments using a modified version of MFCC, called log frequency power coefficients (LFPC) [5]. Pitch contour, intonation, and speech style are categorized as supra-frame features. Bänziger explicitly investigated the role of pitch in emotional speech [6]. They argued that statistics of pitch information, such as F0 mean or F0 range, convey considerable information about emotional status while the shape of F0 contour includes little amount of information. Results using pitch information in emotion recognition systems can be found in many prior efforts [4][7][8]. Studies have shown that higher level linguistic information such as lexical and discourse features can be useful for emotion recognition. In his analysis on segmental features, Lee proposed the quantitative measurement of speakers' attitudes contained in each word using mutual information as a segmental feature [9]. Although the lexical features are beyond the scope of this work, it motivated the combination of acoustic, lexical, and discourse information for emotion recognition.

In the proposed real-time emotion detection system using multi-modal fusion approach of intra- and supra- frame level features, we introduce a new information fusion algorithm which takes care of the characteristics of individual machine algorithms associated with different timescale features. In this framework, the system can be connected to automatic speech recognizer (ASR) systems in distributed computing environment to retrieve additional information, such as lexical or discourse features.

II. System Description

The system consists of four major parts; speech acquisition, feature extraction at each timescale level, machine learning for each feature set, and information fusion to merge the information. Fig. 1 illustrates the basic concept of the system. Note that the speech acquisition process is running simultaneously with other processes using multi-threaded program technology which enables the system to work in real-time. All procedures are programmed in C++ language along with some popular libraries, such as IT++ [10] and TORCH [11].

In real-time applications, however, determining explicit boundaries of phone or utterance is very difficult and requires high computing power. Therefore, in this work, the speech acquisition process provides with fixed length of speech segments. When the fixed length of speech segment is fed into the system through the speech acquisition process, it is split into spectral analysis component and prosody analysis component, which extract the intra- and supra-frame level information respectively. After computing likelihoods (or probabilities) separately with corresponding machine learning algorithms, information fusion is performed to make a decision. These procedures, i.e. feature extraction, computing likelihood, and information fusion, should be done by the time the next speech segment is ready through the speech acquisition procedure.

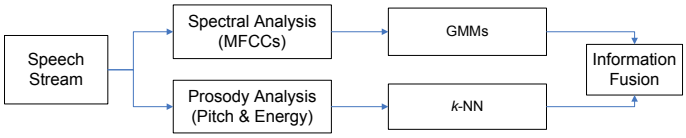

Fig. 1. A simple diagram of the emotion detection system.

A. Intra-Frame Feature: Spectral Features

Mel frequency cepstral coefficient (MFCC) is the most widely used spectral representation of the speech signal in many applications, such as speech recognition and speaker recognition. It is usually extracted in 20-30 ms segments of speech with 50% of overlap. Consequently, approximately 100 feature vectors are generated for every second of signal stream. A learning algorithm needs to be employed for operating on such a large number of feature vectors.

In this work, the Gaussian mixture model (GMM) [12] is adopted to represent the distribution of the features. Under the assumption that the feature vector sequence $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ is an independent identical distribution (i.i.d) sequence, the estimated distribution of the D-dimensional feature vector $\mathbf{x}$ is a weighted sum of M component Gaussian densities $N_i(\mathbf{x})$, each parameterized by a mean vector μ_i and covariance matrix $\mathbf{K}_i$; the mixture density for the model $\mathbf{\Lambda_m}$ is defined as

$$f\left(\mathbf{x}|\mathbf{\Lambda_m}\right) = \sum\nolimits_{i=1}^{M} p_i N_i(\mathbf{x}). \tag{1}$$

The mixture weight p_i satisfies the constraint $\sum_{i=1}^{M} p_i = 1$. We use the expectation maximization (EM) algorithm for the mixtures to get maximum likelihood [12][15] with twelfth order of MFCC.

B. Supra-Frame Feature: Prosody Features

In this work, we utilized the statistics of pitch and energy as prosodic features. We extract the pitch and energy contours for a given segment, and calculate their statistics to construct a twelfth order feature vector. The statistics used are mean, standard deviation, maximum, minimum, median, and jitter. In constrast to the intra-frame feature, only one feature vector is generated per utterance. In this paper, the k nearest neighborhood algorithm (k-NN) [15] is chosen to model the prosodic features. It is a simple and non-parametric machine learning algorithm, which classifies the input based on the prototypes in the training data.

The posteriori probability that given a feature vector $\mathbf{x}$ belongs to class m is

$$P\left(\Lambda_m|\mathbf{x}\right) = P_m\left(\mathbf{x}\right) = \frac{N_m}{k} \quad , \tag{2}$$

where N_m denotes the number of prototypes which belong to the class m among the k nearest prototypes. Since k-NN is based on Euclidean distance, we normalize each component of prosody feature in the training data so that it has zero mean and unit standard deviation.

C. Information Fusion Algorithm

Starting with a simple binary classification theory, a classifier function $C(\cdot)$ which takes a input speech $\mathbf{s}$ will yield a result as to which hypothesis the speech belongs based on a threshold comparison, i.e,

$$C\left(\mathbf{s}\right) = \begin{cases} H_0 & ;\xi \geq \theta \\ H_1 & ;\xi < \theta \end{cases} , \tag{3}$$

where

$$\xi = S_0 - S_1 \tag{4}$$

and S_m represents likelihood that the speech $\mathbf{s}$ belongs to H_m. In this work, we set the hypothesis as

- H_0: the input speech is of angry emotional status
- H_1: the input speech is of neutral emotional status.

The decision arising from the spectral and prosodic feature classifiers need to be combined in order to have a unique and more accurate classification. Many algorithms have been proposed to deal with multiple modalities. One of the most simple and popular methods is a weighted sum of likelihoods from different modalities with a weighing factor that is empirically determined. In the case of two modalities, it can be represented as

$$\begin{aligned} S_m &= \lambda \cdot LL\left(\mathbf{s}|\Lambda_m^{M_1}\right) + (1-\lambda) \cdot LL\left(\mathbf{s}|\Lambda_m^{M_2}\right) \\ &= \log\left(P\left(\mathbf{s}|\Lambda_m^{M_1}\right)\right)^{\lambda} + \log\left(P\left(\mathbf{s}|\Lambda_m^{M_2}\right)\right)^{1-\lambda}, \end{aligned} \tag{5}$$

where λ and $\Lambda_m^{M_i}$ denote weighting value and model of i-th modality for H_m, respectively. We set M_1 for prosody features and M_2 for spectral features. Consequently, the final score is

$$\begin{aligned} \xi &= S_0 - S_1 \\ &= \log\left(\frac{P\left(\mathbf{s}|\Lambda_0^{M_1}\right)}{P\left(\mathbf{s}|\Lambda_1^{M_1}\right)}\right)^{\lambda} + \log\left(\frac{P\left(\mathbf{s}|\Lambda_0^{M_2}\right)}{P\left(\mathbf{s}|\Lambda_1^{M_2}\right)}\right)^{1-\lambda} \\ &= \log\left(\frac{P\left(\Lambda_0^{M_1}|\mathbf{s}\right)P\left(\Lambda_0^{M1}\right)}{P\left(\Lambda_1^{M_1}|\mathbf{s}\right)P\left(\Lambda_1^{M_1}\right)}\right)^{\lambda} + \log\left(\frac{P\left(\mathbf{s}|\Lambda_0^{M_2}\right)}{P\left(\mathbf{s}|\Lambda_1^{M_2}\right)}\right)^{1-\lambda} \\ &= \log\left(\frac{P\left(\Lambda_0^{M_1}|\mathbf{s}\right)}{P\left(\Lambda_1^{M_1}|\mathbf{s}\right)}\right)^{\lambda} + \log\left(\frac{P\left(\mathbf{s}|\Lambda_0^{M_2}\right)}{P\left(\mathbf{s}|\Lambda_1^{M_2}\right)}\right)^{1-\lambda} + C \end{aligned} \tag{6}$$

where

$$C = \lambda \cdot \log\left(P\left(\Lambda_0^{M_1}\right)/P\left(\Lambda_1^{M_1}\right)\right) \tag{7}$$

and it is from the difference between the priori probabilities of the classes: $C = 0$ when $P\left(\Lambda_0^{M_1}\right) = P\left(\Lambda_1^{M_1}\right)$. Since C only affects the threshold θ in binary decision tasks, the posteriori probability from k-NN can be used instead of likelihood without changing the performance.

This algorithm, however, needs to be modified for this work due to the nature of two different learning algorithms. While GMM yields very small positive likelihood value from probability density function, the k-NN model yields the discrete posteriori probability value with $1/k$ resolution from 0 to 1, and in the limit case of $P\left(\Lambda_m|\mathbf{s}\right) = 0$ which is computationally impossible to calculate the logarithm. Although adding very small positive value ϵ (10^{-50} in this work) might help the computation, the decision would be dominated by the result and to find appropriate weighting factor would be difficult.

Therefore, we propose a modification on the weighted sum of likelihood, by using posteriori probability instead of the log probability for the k-NN model, i.e

$$\begin{aligned} S_m &= \lambda \cdot P\left(\Lambda_m^{M_1}|\mathbf{s}\right) + (1-\lambda) \cdot LL\left(\mathbf{s}|\Lambda_m^{M_2}\right) \\ &= \log\left(e^{P_m(\mathbf{s})}\right)^{\lambda} + \log\left(P\left(\mathbf{s}|\Lambda_m\right)\right)^{1-\lambda} \end{aligned} \tag{8}$$

Then, the final score is represented as

$$\begin{aligned} \xi &= S_0 - S_1 \\ &= \log\left(\frac{e^{P_0(\mathbf{s})}}{e^{P_1(\mathbf{s})}}\right)^{\lambda} + \log\left(\frac{P\left(\mathbf{s}|\Lambda_0\right)}{P\left(\mathbf{s}|\Lambda_1\right)}\right)^{1-\lambda}. \end{aligned} \tag{9}$$

Briefly speaking, the proposed fusion algorithm provides the same impact with simply taking exponential of the posteriori probability from k-NN models inside the logarithm. It prevents zero or infinity inside to compute logarithm, and also helps to find the optimal or sub-optimal weighting factor for the task by normalizing the dynamic range of likelihood.

III. Experimental Setups

A. Database

The EMA database is used for model training. It was collected for emotional speech related research at USC. The train set has 3 speakers (1 male and 2 female) and 10 types of sentences. Each speaker reads the sentences 5 times, and the sentences are approximately two seconds long. The recording was repeated 4 times for each emotion. For testing, data from a similar collection but with different subjects is used. There are 4 subjects (1 male and 3 female) in the test database. The sampling rate of the speech signal is 16kHz and stored in 16-bit resolution. See [13] for details about the database. For this particular work, we only use neutral and angry sessions to formulate a binary decision problem (200 utterances for train and 1,764 utterances for test).

B. Evaluation

Since evaluating the performance of the real-time system is very difficult, we concatenate the utterances from the test database into fixed length segments and feed them into the system to simulate real-time conditions. A toolkit, called DETware from NIST, is used to evaluate the system performance [14]. Although the toolkit is originally devised for speaker verification systems, it can be generally used for any binary decision problems by showing detection error trade-off depending on the threshold. Equal error rate (EER) is usually the quoted metric for representing the quality of the detection, and it represents the error rate when the false alarm and the miss probability are equal.

IV. Results and Discussions

Table I shows the equal error rates (EER) of each single modality task, spectral and prosody features, with respect to various model settings and length of speech segment. For the spectral feature cases, an improvement is generally observed as the speech segment gets longer, while it is not necessarily always true for the prosody feature cases. It is reasonable because the more spectral feature vectors there are the longer the segment is, while there is only one feature vector for the prosody feature regardless of the length of segment. For prosody features, the EER is lower when k is large for short segments (< 3 sec.), while it is minimized at $k = 10$ for longer segments (> 3 sec.). In the case of spectral features, the EER is minimized when the number of Gaussian mixture is 64 for all segment lengths. It indicates k should be adjusted according to the length of the segment in real applications, while the number of GMM can be fixed regardless of the length. Fig. 2 and 3 depict DET curves of the chosen tasks in Table I. The curves in Fig. 3 are in staircase-like shape because the scores from k-NN are discrete so that threshold and EER corresponding the threshold are also discrete.

TABLE I

EER for various speech length and (upper) number of mixtures for GMM with spectral feature, (bottom) k for k-NN model with prosody features.

speech length (sec)		1 sec	2 sec	3 sec	4 sec	5 sec
Spectral features	$M = 8$	24.4	20.0	18.8	17.9	15.5
	$M = 16$	24.3	19.4	17.9	17.1	16.3
	$M = 32$	24.4	19.3	18.2	16.8	16.8
	$M = 64$	**22.1**	**17.9**	**15.7**	**15.2**	**14.5**
	$M = 128$	23.7	18.9	17.4	16.0	15.9
Prosody features	$k = 1$	31.1	18.8	17.9	20.1	21.8
	$k = 5$	21.6	16.4	15.1	10.5	10.3
	$k = 10$	21.5	15.3	12.3	**8.23**	**8.10**
	$k = 20$	**19.3**	14.3	11.8	9.66	10.3
	$k = 30$	19.7	**13.8**	**11.4**	9.34	9.90

Table II provides comparison between the proposed and conventional algorithm in terms of EER for various settings. Weighting value λ for minimizing EER and its corresponding EER are presented. Note that either fusion algorithm outperforms the individual modalities, which indicates that the information included in each level of feature is complementary. It is remarkable that the proposed algorithm outperforms the conventional one. For example in $k = 5$ and $k = 10$ cases for 5 seconds long segments, we obtain 33% and 27.3% of relative improvement over the conventional algorithm. One might notice that the weighting value λ in parenthesis, which minimizes the EER with a given setting, varies depending on the segment length: greater λ is required for the longer segments. It indicates that λ also needs to be adjusted according to the length of the segment in real applications.

Fig. 4 and 5 illustrate the performance improvement due to the proposed multi-modal approach for chosen model setting cases ($k =$

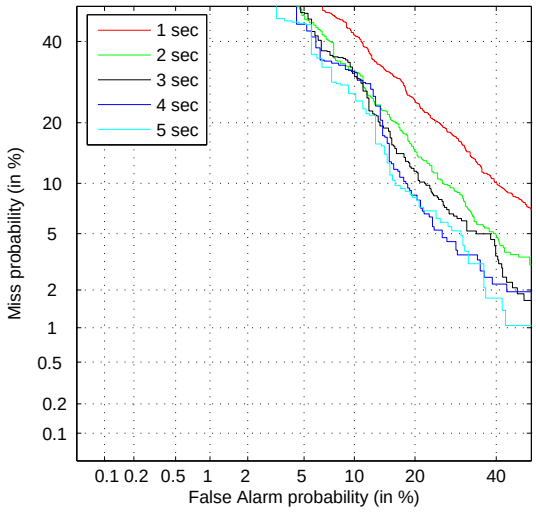

Fig. 2. DET curves with various speech length for spectral model with 64 mixtures.

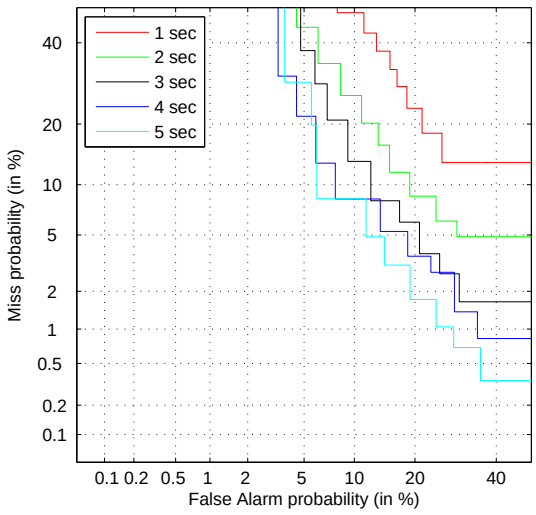

Fig. 3. DET curves with various speech length for prosody model with 10-NN model.

TABLE II

MINIMUM EER WITH VARIOUS SPEECH LENGTH AND k ($M = 64$), AND THE CORRESPONDING WEIGHTING VALUE λ IN THE PARENTHESES.

		1 sec	2 sec	3 sec	4 sec	5 sec
$k = 1$	Proposed	18.5 (0.47)	12.9 (0.44)	11.4 (0.62)	10.1 (0.67)	8.91 (0.66)
	Conventional	18.9 (0.02)	12.9 (0.02)	11.4 (0.04)	10.1 (0.05)	8.91 (0.05)
$k = 5$	Proposed	17.1 (0.57)	11.5 (0.66)	9.03 (0.72)	8.23 (0.74)	5.32 (0.72)
	Conventional	17.2 (0.04)	13.0 (0.05)	10.6 (0.74)	8.23 (0.69)	7.93 (0.64)
$k = 10$	Proposed	16.9 (0.57)	10.6 (0.66)	9.03 (0.73)	7.44 (0.78)	4.75 (0.78)
	Conventional	17.7 (0.37)	12.5 (0.53)	10.4 (0.85)	7.91 (0.65)	6.54 (0.65)
$k = 20$	Proposed	16.1 (0.63)	10.8 (0.70)	8.68 (0.73)	7.44 (0.78)	5.15 (0.78)
	Conventional	15.7 (0.36)	11.4 (0.48)	9.27 (0.65)	7.58 (0.78)	5.15 (0.64)
$k = 30$	Proposed	15.8 (0.63)	10.3 (0.72)	8.20 (0.72)	6.80 (0.88)	5.73 (0.75)
	Conventional	15.8 (0.42)	10.6 (0.52)	8.93 (0.69)	6.80 (0.79)	6.13 (0.66)

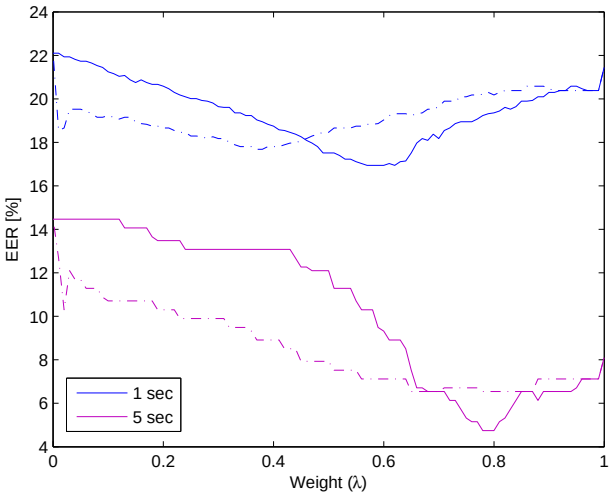

Fig. 4. EER versus weighting value λ ($k = 10$ and $M = 64$). Solid and dashed lines for proposed and conventional algorithm respectively.

10 and $M = 64$). For simplicity, only one and five seconds long segment cases are chosen. In both figures, solid lines and dashed lines represent the output of the proposed algorithm and the conventional algorithm, respectively. Fig. 4 depicts that EER versus weighting value λ. For conventional algorithm, one can observe that there is a steep EER drop in small λ but does not reach the minimum EER which the proposed algorithm accomplishes. We argue that this is because of the characteristics of k-NN which described earlier in Section II-C and the performance improvement is from using the posteriori probability itself instead of taking logarithm. The DET curves which correspond to the minimum EER are depicted in Fig. 5. It verifies that the proposed algorithm is also superior than the conventional one in terms of DET curves.

V. CONCLUSIONS AND FUTURE WORK

A real-time emotion detection system using a new information fusion algorithm of intra- and supra- frame features was developed. Experimental results to verify the real-time system showed that the

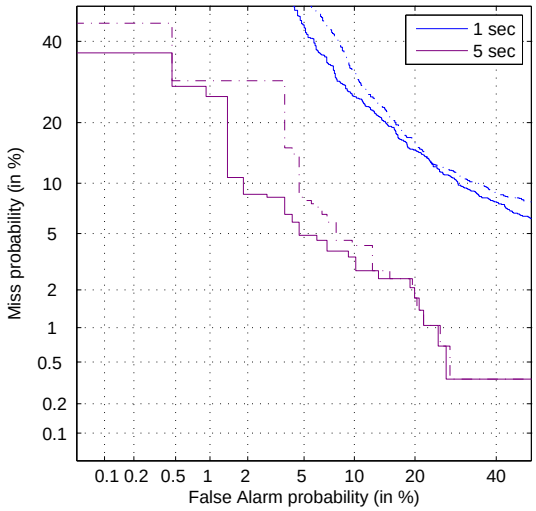

Fig. 5. DET curves which minimizes the EER for given settings ($k = 10$ and $M = 64$).

proposed system with multi-modal fusion algorithm outperforms the individual modalities, and further improvements were obtained by using the proposed information fusion algorithm over the conventional weighted sum of likelihood fusion algorithm in terms of equal error rate (EER) as well as detection error trade-off (DET) curves. The proposed real-time system will be extended to be connected to automatic speech recognizer (ASR) systems, question-answer classifier, and etc. in distributed computing environment to retrieve additional information.

REFERENCES

[1] E. Leon, G. Clarke, and V. Callaghan, "Towards a robust real-time emotion detection system for intelligent buildings," *IEE International Workshop, Intelligent Environments 05*, pp. 162- 167, 2005.

[2] I. R. Murray, J.L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *Journal of Acoustical Society of America*, Vol. 93, pp. 1097-1108, 1993.

[3] M. Airas, P. Alku, "Emotions in short vowel segments: effects of the glottal flow as reflective by the Normalized Amplitude Quotient (NAQ)," *ADS 2004*, pp. 13-24, 2004.

[4] O. W. Kwon, K Chan, J. Hao, T. -W. Lee, "Emotion recognition by speech signals," *In Proc. of EUROSPEECH 2003*, 2003.

[5] T. L. Nwe, S.W. Foo, L. C. Silva, "Speech emotion recognition using hidden markov models," *Speech Communication*, Vol. 41, pp. 603-623, 2003.

[6] T. Bänziger, K. R. Scherer, "The role of intonation in emotional expression", *Speech Communication*, Vol. 46, 252-267, 2005.

[7] A. Nogueiras *et al.*,"Speech emotion recognition using hmm," *Proceedings of Interspeech 2001*, 2001.

[8] D. Ververidis, C. Kotropoulos, I. Pitas, "Automatic emotional speech classification," *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processings*, pp. 593-596, 2004.

[9] C. M. Lee and S. Narayanan, "Towards detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13(2), pp. 293-302, 2004.

[10] *http://itpp.sourceforge.com/*

[11] *http://www.torch.ch/*

[12] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture models", *Speech Communication*, vol. 17, pp. 91-108, 1995.

[13] S. Lee, S. Yildirim, A. Kazemzadeh, S. Narayanan, "An articulatory study of emotional speech production," *Proc. of EUROSPEECH 2005*, 2005.

[14] *http://www.nist.gov/speech/tools/*

[15] R. Duda, P. Hart, D. Stock, *Pattern Classification*, second edition, Wiley Interscience, New York, 2000.