Model-Based Parametric Features for Emotion Recognition from Speech[+]

Sankaranarayanan Ananthakrishnan [#1], Aravind Namandi Vembu [*2] and Rohit Prasad [#1]

[#] *Speech, Language and Multimedia Technologies*
Raytheon BBN Technologies
10 Moulton Street
Cambridge, MA 02138, U.S.A.
[1] {sanantha,rprasad}@bbn.com

[*] *Ming-Hsieh Department of Electrical Engineering*
University of Southern California
Los Angeles, CA 90089, U.S.A.
[2] namandiv@usc.edu

***Abstract*—Automatic emotion recognition from speech is desirable in many applications relying on spoken language processing. Telephone-based customer service systems, psychological healthcare initiatives, and virtual training modules are examples of real-world applications that would significantly benefit from such capability. Traditional utterance-level emotion recognition relies on a global feature set obtained by computing various statistics from raw segmental and supra-segmental measurements, including fundamental frequency (F0), energy, and MFCCs. In this paper, we propose a novel, model-based parametric feature set that better discriminates between the competing emotion classes. Our approach relaxes modeling assumptions associated with using global statistics (e.g. mean, standard deviation, etc.) of traditional segment-level features for classification, and results in significant improvements over the state-of-the-art in 7-way emotion classification accuracy on the standard, freely-available Berlin Emotional Speech Corpus. These improvements are consistent even in a reduced feature space obtained by Fisher's Multiple Linear Discriminant Analysis, demonstrating the signficantly higher discriminative power of the proposed feature set.**

I. Introduction

Spoken language contains a wealth of paralinguistic cues that convey information beyond mere lexical import. These are expressed through subtle adjustments to any or all of the following characteristics of speech: pitch, loudness, and intonation. These are often employed in a supplementary role where they enhance the listeners' understanding of spoken words. Pitch accents and boundary tones, for instance, serve as expressions of syllabic stress and phrase breaks, respectively, and play an important role in disambiguation. In other cases, they serve as indicators of the speaker's cognitive/psychological state and attitude. Emotions fall within this latter category. Accurate knowledge of a subject's emotional state is important for successful human-computer interactions. For instance, telephone-based automated customer service systems could, upon detection of frustration or anger, transfer the user to a human representative. Accurate detection of repeated patterns of depression and/or sadness in telephone conversations with health-care providers could facilitate early diagnosis of mental health issues [1]. Emotion recognition is also a key component in virtual training systems [2], where avatars must be influenced by and respond to the trainee's cognitive state.

The availability of suitably annotated corpora in recent years has spurred significant research in automatic emotion recognition from speech [3], [4], [5]. All of these approaches involve extracting numerous segmental and supra-segmental features from the speech signal. Segmental features are time-varying and include short-term spectral measurements such as mel-frequency cepstral coefficients (MFCC) and formants. Supra-segmental features, such as fundamental frequency (F0) and degree of voicing, either vary very slowly or are inherently utterance-level characteristics.

The use of features over different time-scales (segmental and supra-segmental) poses some interesting challenges. First, the number of segmental feature vectors obtainable from a given utterance depends on its length. This forces maximum-likelihood classification using generative time-series models such as Hidden-Markov Models (HMMs), which, while useful for sequence segmentation, do not possess the separation ability of discriminative classifiers such as multi-layer perceptrons (MLPs) or support vector machines (SVMs). Second, there is no well-established method for combining features at different time-scales beyond ad-hoc feature concatenation; for instance, by appending utterance-level features to each segmental feature vector.

Numerous methods have been proposed in the emotion recognition literature for applying discriminative classifiers to features consisting of both segmental and supra-segmental evidence. One is to augment each segment-level feature vector

[+] Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

[2] This research was conducted during the second author's internship at Raytheon BBN Technologies.

with global utterance-level features, and perform emotion classification at the frame level. Weighted majority voting over all segments can then be used to classify the utterance. However, this approach does not offer optimal classification in the mathematical sense. Another solution is to compute global statistics from segment-level features and combine these with the supra-segmental cues. Given a sequence of MFCC vectors, for instance, we can compute their sample mean and augment it with other global features such as average F0. The disadvantage of this approach is that, by averaging, we greatly reduce the effective dimensionality of segmental features. In the process, we discard information potentially useful for classification. Finally, a relevant approach is that of Vlasenko et al. [6], [7], who proposed a two-level approach where frame-level scores from a Gaussian Mixture Model (GMM) are integrated with global utterance-level features and fed into a turn-level SVM emotion classifier.

In this paper, we propose a novel, model-based parametric approach to utterance-level feature extraction from a sequence of segment-level features. Rather than average them across all frames or use scores from frame-level models for utterance-level classification, we use the segment-level features to determine the location of the utterance in a *model space*, and use weighted averages of the model parameters as input features for a discriminative classifier. Under this framework, we are able to significantly relax modeling assumptions associated with the ad-hoc feature averaging approach. Using simple generative probability models, we are also able to obtain features of much higher dimensionality than feature averaging, thereby preserving more discriminative information. The proposed method is fairly general and can be applied to many classification problems where local and global features must be merged. Further, any suitable parametric model of time-series data can be used to define the model space. We use Gaussian Mixture Models (GMMs) in our work.

The remainder of this paper is organized as follows. Section II describes the emotion-labeled speech corpus used in our work. Section III summarizes the different types of basic acoustic features we use for emotion classification. Section IV describes how we obtain a high-dimensional model-based parametric feature vector from a variable-length collection of segment-level feature vectors. Experimental details and results are summarized in Section V. Section VI concludes the paper with a brief summary of our work and presents future directions for research.

II. Data Corpus

The Berlin Emotional Speech Corpus [8] is a freely-available data set, which has in recent years become a standard for emotion recognition research. This German-language corpus consists of 535 short utterances spoken by ten native speakers in seven different emotions (anger, boredom, disgust, anxiety/fear, happiness, sadness, and neutral). One of ten semantically neutral sentences constitutes the content of each spoken utterance. The corpus is gender balanced with equal number of male and female speakers. Each spoken utterance corresponds to exactly one emotion type, i.e. emotion labels are assigned at the utterance level. Tables I and II summarize the distribution of various emotion categories across speakers and content, respectively.

TABLE I
DISTRIBUTION OF EMOTIONAL UTTERANCES ACROSS SPEAKERS.

Emo/Spk	*03*	*08*	*09*	*10*	*11*	*12*	*13*	*14*	*15*	*16*
Anger	14	12	13	10	11	12	12	16	13	14
Boredom	5	10	4	8	8	5	10	8	9	14
Disgust	1	0	8	1	2	2	8	8	5	11
Anxiety	4	6	1	8	10	6	7	12	8	7
Happiness	7	11	4	4	8	2	10	8	6	11
Sadness	7	9	4	3	7	4	5	10	4	9
Neutral	11	10	9	4	9	4	9	7	11	5
Total	49	58	43	38	55	35	61	69	56	71

TABLE II
DISTRIBUTION OF EMOTIONAL UTTERANCES ACROSS SENTENCES.

Emo/Sent	*a*					*b*				
	01	*02*	*04*	*05*	*07*	*01*	*02*	*03*	*09*	*10*
Anger	12	14	14	15	10	11	15	11	12	13
Boredom	6	8	7	9	11	8	6	9	9	8
Disgust	6	9	2	4	4	7	2	4	4	4
Anxiety	7	7	7	8	6	6	7	4	8	9
Happiness	7	6	8	8	9	9	6	7	7	4
Sadness	1	7	6	10	8	3	6	10	6	5
Neutral	10	9	7	8	7	7	9	7	6	9
Total	49	60	51	62	55	51	51	52	52	52

Each utterance in the corpus is labeled with its emotion category, as well as speaker and content identity. Since the corpus is relatively small by machine learning standards, researchers using this corpus have generally opted for cross-validation style performance evaluation. Popular configurations in the literature include *leave-one-text-out* and *leave-one-subject-out* evaluation, both of which are presented in this paper. These configurations allow us to directly compare our system to previously published performance evaluations [5].

III. Basic Acoustic Features

We use numerous features derived from F0, voicing, jitter, shimmer, intensity (loudness) and formants, as well as segmental MFCC features, which are standard in speech recognition. Following is a summary of various feature types used in this paper, all of which were extracted from the speech signals using Praat [9], a widely used, multi-platform, open-source phonetics program for manipulating and analyzing speech. We refer to the first five feature groups as U-FEAT (utterance-level features).

A. Fundamental frequency (F0)

Pitch or fundamental frequency (F0) has been shown in previous work [4], [5], [3] to be a useful feature for emotion recognition. We used a pitch extraction algorithm based on the autocorrelation method as implemented in Praat. As pitch-tracking is susceptible to halving and doubling errors, we

performed further post-processing using a five-point median filter, followed by explicitly removal of points significantly divergent from average F0. From the processed F0 track, we extracted three utterance-level features including the range, mean, and standard deviation.

B. Formants (F1, F2)

The effect of emotion on formants has also been shown and observed [10], [11], providing the case for inclusion of formant based features. In particular, the first and second formants of speech are known to be influenced by emotion [11] and are expected to provide useful information for emotion classification. The first and second formants were extracted from the speech signals based on the Burg method [12] to compute LPC coefficients using a window length of 25ms. Similar to F0, utterance-level statistics obtained from the formant contours (range, mean, and standard deviation) are used as features for emotion recognition. A total of six formant-related features are used.

C. Intensity

Certain emotions such as anger and happiness are often correlated with bursts of loudness in speech. To account for this, we extract intensity, an indicator of loudness, from the speech signals. In Praat, this feature is measured in dB SPL (relative to 2×10^{-5} Pascal). We compute the maximum, mean and standard deviation of intensity across the entire utterance for a total of three utterance-level intensity features.

D. Jitter and shimmer

Jitter is a voice quality feature that refers to the variation in pitch, which can cause a "rough" sound. It is generally defined as the absolute or relative difference between consecutive periods in a segment of voiced speech. Studies have shown that jitter is a correlate of negative emotions such as sadness/depression [1]. We use Praat to compute local, absolute, and perturbation values for a total of five jitter features averaged across the utterance. Shimmer is the analogous voice quality measure that evaluates the degree of local change in the amplitude or intensity of speech. It is usually a function of the difference between amplitudes of consecutive periods of a voiced speech segment. Again, we used Praat to obtain a total of six utterance-level shimmer features obtained by averaging local shimmer values.

E. Voicing statistics

Voice breaks can occur frequently during excited emotions (either positive or negative) and have the potential to discriminate between certain types of emotions. We compute the fraction of locally unvoiced frames for an utterance, as well as the number of voice breaks (number of inter-pulse intervals longer than a certain threshold). We also evaluate the degree of voice breaks, defined as the ratio of total duration of the breaks between voiced segments of speech to the total duration of the speech signal. These yield three inherently utterance-level features, eliminating the need to compute averages or other statistics.

F. Segmental MFCC

Bozkurt et al. [13] showed that standard MFCC features yield surprisingly good performance on the emotion recognition task. We used Praat to extract 12 MFCC features over 30ms windows (frame rate = 100 Hz) along with the five-point delta and acceleration values to obtain a 36-dimensional feature vector for each frame. In our baseline system, we convert segmental MFCC feature vectors to a single utterance-level feature by computing the range, mean and standard deviation of each dimension. This yielded basic utterance-level MFCC features (B-MFCC) of 108 dimensions.

IV. Model-based Parametric Features

As discussed in Section I, using a discriminative classifier (logistic regression, MLP, SVM, etc.) in conjunction with segmental and global features is non-trivial. Frame-level feature augmentation and classification followed by weighted majority voting is one ad-hoc solution, but it does not guarantee mathematically optimal utterance-level classification. At the other end of the spectrum is the widely-used practice of computing global statistics over the segmental features (e.g. mean, median, standard deviation, etc.). While this is acceptable for slowly varying measurements such as F0, its appropriateness for segmental features, whose trajectory can be highly non-stationary, is questionable. This method forces a significant reduction in the effective dimensionality of segmental features, which can cause degradation in classification accuracy.

To alleviate this problem, we introduce the idea of a *model space* defined by a class-conditional parametric representation of segmental features. In this paper, we assume a generative, probabilistic form for the latter. Assuming M class labels, let $\mathbf{\Theta}_1, \ldots, \mathbf{\Theta}_M$ represent parameter vectors of class-conditional probability distributions over the corresponding sets of segmental features. If the underlying models are P-mixture GMMs, for instance, $\mathbf{\Theta}_i = \{\alpha_i^{1\ldots P}, \mu_i^{1\ldots P}, \mathbf{\Sigma}_i^{1\ldots P}\}$ is the concatenation of the mixture weights, mean vectors, and covariance matrices of the probability model representing the i^{th} class. Regardless of the actual form of the models, we assume that they support the evaluation of class-conditional likelihoods $p_i(\mathbf{x} \mid \mathbf{\Theta}_i)$ for arbitrary segmental feature vector $\mathbf{x}$. In the case of GMMs, this is simply the likelihood of the observation x given the GMM parameters (mixture weights, mean vectors, and covariance matrices).

Based on the above description, we can compute the normalized likelihood of segmental feature vector $\mathbf{x}$ with respect to each class-conditional model as shown in Equation 1.

$$\lambda_k(\mathbf{x}) = \frac{p_k(\mathbf{x} \mid \mathbf{\Theta}_k)}{\sum_{i=1}^{M} p_i(\mathbf{x} \mid \mathbf{\Theta}_i)} \tag{1}$$

Under the interpretation that $\lambda_k(\mathbf{x})$ is the probability of $\mathbf{x}$ belonging to the k^{th} class, we can compute its location in model space as the weighted average of the corresponding parameters. This allows us to transform the segmental feature vector $\mathbf{x}$ to a potentially much higher dimension, depending on the complexity of the underlying models. Finally, we average transformed feature vectors over the entire utterance to obtain

a representative feature vector of fixed dimensionality. This is illustrated in Equation 2.

$$\begin{aligned}\mathbf{y} &= f(\mathbf{x}_1, \ldots, \mathbf{x}_T) \\ &= \frac{1}{T}\sum_{j=1}^{T}\sum_{i=1}^{M}\lambda_i(\mathbf{x}_j)\boldsymbol{\Theta}_i \\ &= \sum_{i=1}^{M}\boldsymbol{\Theta}_i\frac{1}{T}\sum_{j=1}^{T}\lambda_i(\mathbf{x}_j) \\ &= \sum_{i=1}^{M}\beta_i\boldsymbol{\Theta}_i \qquad (2)\end{aligned}$$

$$\text{where } \beta_i = \frac{1}{T}\sum_{j=1}^{T}\lambda_i(\mathbf{x}_j)$$

The above mapping from a sequence of low-dimensional segmental feature vectors $\mathbf{x}_1, \ldots, \mathbf{x}_T$ to a single feature vector $\mathbf{y}$ of fixed but potentially much higher dimensionality has many benefits. First, it relaxes modeling assumptions associated with simple global statistics of the feature sequence. For instance, when the underlying class-conditional models are 5-mixture, full-covariance GMMs, a (sequence of) 36-dimensional MFCC vector(s) is represented by a 6665-dimensional vector in model space (1 mixture weight, 36 mean components, and 1,296 covariance components for each of the 5 mixtures). On the other hand, feature averaging is equivalent to modeling segmental features of each utterance using a separate single-mixture Gaussian with a constant covariance matrix. Under this interpretation, the proposed features derived from class-conditional models should intuitively possess more discriminative power than global statistics of segmental features, which do not rely on class labels.

Second, $\mathbf{y}$ can be interpreted as a linear combination of M "basis vectors" estimated from the labeled data. While individual segmental feature vectors from different parts of the same utterance may exhibit large differences, their projections in model space are likely to exhibit much lower variance due to the effect of aggregation in the class-conditional models. Reduced within-class scatter could potentially yield better classification performance. This aspect of parametric features is similar to codebook learning, where an arbitrary feature vector is expressed as a combination of representative codewords (typically obtained by k-means clustering or vector quantization).

The choice of underlying class-conditional models depends on the application and available training data. HMMs and GMMs present two natural choices in our current work on emotion recognition. We use GMMs due to their robustness and relative ease of estimation from limited data. Segmental feature vectors (MFCCs) belonging to each emotion class across all utterances in the training partition are pooled and used to estimate the parameters of the corresponding GMM using the well-known expectation-maximization algorithm. Model-space features are then derived according to Equation 2 for each training/testing utterance, and are input to a discriminative classifier for training/evaluation as described in Section V. We refer to these features as P-MFCC (parametric model-based MFCC). We used the open-source Netlab toolbox [14] for extracting these parametric features from raw segmental MFCCs, and their deltas and accelerations.

V. Experimental Results

Given the small size of the emotion labeled corpus, we performed two types of cross-validation for reliable performance evaluation. The first is *leave-one-text-out* partitioning, in which utterances corresponding to each text are successively held-out for evaluating classification performance. In each case, the remaining utterances are used to train a discriminative classifier. The other cross-validation method is *leave-one-subject-out*, in which utterances corresponding to each speaker are held-out instead. Since there are ten unique texts and speakers, both approaches effectively perform ten-fold cross-validation on different subsets. In both cases, we measure *weighted average accuracy* A_w, defined as the ratio of the total number of correctly labeled utterances to the total number of utterances across all held-out test sets; and *unweighted average accuracy* A_u, which is the average classification accuracy of each emotion class across all held-out sets. This follows the approach of Tawari [5], and enables us to directly compare emotion classification accuracy to the state-of-the-art.

Weka [15], an open-source machine learning toolkit, offers easy access to a variety of discriminative classifiers. For classification experiments with full feature sets, we use a support vector machine (SVM) that implements the sequential minimal optimization (SMO) algorithm [16] using a polynomial kernel with unit exponent. This classifier has been shown to avoid overfitting and performs well when the dimensionality of the feature space is large. We also evaluate emotion labeling accuracy in a reduced feature space obtained by applying Fisher's Linear Discriminant Analysis to the feature sets. We use a multi-layer perceptron (MLP) [17] with ten hidden nodes for emotion classification in this low-dimensional feature space.

A. Number of GMM Mixtures for P-MFCC Features

The number of GMM mixtures used to project segmental MFCCs to a high-dimensional feature vector is a design parameter. We expect that very few mixtures may not be sufficient to capture the information contained in the MFCC features. On the other hand, increasing the number of mixtures beyond a point could cause the curse of dimensionality to overcome the increase in captured information and lead to poor performance. In order to investigate the effect of this parameter, we performed classification experiments for different number of mixtures, as shown in Table III. We used the SMO/SVM classifier for this task. Although ideally the choice of number of mixtures should be optimized on a held-out development set, none was available in this case because we wanted to replicate the cross-validation configuration of Tawari [5]. As expected, accuracy initially increased with number of mixtures and reached its peak with five mixtures

before declining. Performance evaluation was based on the *leave-one-text-out* cross-validation configuration. Based on this experiment, we set the number of mixtures for class-conditional GMMs to five for computing P-MFCC features.

TABLE III
GMM MIXTURES VS. CLASSIFICATION ACCURACY FOR P-MFCC

Mixtures	A_u	A_w
1	76.5%	78.3%
3	80.3%	81.3%
5	**83.2%**	**83.9%**
7	82.0%	83.4%

Note that full-covariance matrices were used in all of the above cases. We found that classification accuracy with parametric features derived from diagonal covariance GMMs was consistently worse, suggesting that the implicit assumption of uncorrelated features does not hold, causing loss of information.

B. Comparing Feature Sets using Fisher's LDA

Since the three types of features, namely U-FEAT, B-MFCC, and P-MFCC vary greatly in their dimensionality, it could be argued that the larger number of classifier parameters used by, say, P-MFCC is what results in improved performance. In order to provide a fair comparison between the feature sets of varying complexity, we use Fisher's Linear Discriminant Analysis (LDA) to reduce all feature sets to the same (low) dimensionality. Fisher's LDA defines a mapping on the feature set that aims to reduce the intra-class scatter, while maximizing the inter-class mean distance. The multi-class version of Fisher's LDA results in a transformation matrix, which projects the full feature vector down to a reduced feature space whose dimensionality is equal to one less than the number of classes (with seven emotion labels, the transformed features were six-dimensional).

We used the LDA implementation of the open-source PRTools toolbox [18] in Matlab. LDA was applied to each of the three feature sets above. MLP classifiers with identical structure and configuration settings (ten hidden nodes, 500 training iterations, backpropagation learning) were used for classifying each set. While we expect slightly lower classification accuracy than using the full set of features (due to the drastic reduction in dimensionality), we expect a consistent trend where the most informative reduced features exhibit the best performance. Ultimately, the goal of this experiment is to be able to compare different feature sets on an equal footing.

For the *leave-one-text-out* configuration, we see in Table IV that the U-FEAT set performs the worst, while P-MFCC features result in highest accuracy. These results clearly show that, even in a greatly reduced feature space, P-MFCC possesses better discriminative power than B-MFCC and U-FEAT. Interestingly, the B-MFCC features alone tend to perform better than U-FEAT. This is consistent with the work of Bozkurt et al. [13] and underlines the importance of MFCC-based features in emotion recognition.

TABLE IV
LEAVE-ONE-TEXT-OUT ACCURACY FOR LDA-REDUCED FEATURES

Feature Set	A_u	A_w
U-FEAT	65.0%	67.10%
B-MFCC	73.7%	74.77%
P-MFCC	**78.4%**	**79.25%**

C. Full Feature-Set Classification

The LDA experiments illustrated that P-MFCC features contain more information about the emotion classes than B-MFCC and U-FEAT. In this experiment, we performed classification experiments using the entire feature set without feature selection or reduction. Since the number of features can range in the hundreds (B-MFCC) or thousands (P-MFCC), we used the SMO/SVM classifier for this task. We considered two additional configurations: (U-FEAT + B-MFCC) and (U-FEAT + P-MFCC), in which utterance-level features U-FEAT were combined with B-MFCC and P-MFCC, respectively. Identical classifier settings were used across all five configurations.

Table V summarizes the weighted and unweighted classification accuracy of these configurations for the *leave-one-text-out* cross-validation method. We note that P-MFCC by itself outperforms combinations of all other feature sets; the (U-FEAT + P-MFCC) configuration gives a slight further improvement. We also note that our best emotion classification (weighted) accuracy of 84.1% on this configuration is a significant improvement over the state-of-the-art performance of 81.4% reported by Tawari [5]. Table VI shows the aggregate confusion matrix for this best-performing configuration.

TABLE V
LEAVE-ONE-TEXT-OUT ACCURACY FOR FULL FEATURE SETS

Feature Set	A_u	A_w
U-FEAT	57.8%	62.1%
B-MFCC	73.6%	75.1%
P-MFCC	83.2%	83.9%
U-FEAT + B-MFCC	79.5%	80.4%
U-FEAT + P-MFCC	**83.4%**	**84.1%**

TABLE VI
LEAVE-ONE-TEXT-OUT CONFUSION MATRIX FOR U-FEAT + P-MFCC

Reference	**Predicted**						
	a	b	c	d	e	f	g
a = Anger	117	0	0	0	10	0	0
b = Boredom	0	70	2	0	0	3	6
c = Disgust	2	2	39	1	2	0	0
d = Anxiety/Fear	3	2	3	52	7	0	2
e = Happiness	19	0	1	3	48	0	0
f = Sadness	0	1	0	0	0	59	2
g = Neutral	0	11	1	2	0	0	65

Additionally, Table VII summarizes emotion classification accuracy on the above feature configurations for the *leave-one-subject-out* cross-validation method. We note that there is a significant drop in performance across all feature sets compared to the *leave-one-text-out* evaluation approach. This

TABLE VII
LEAVE-ONE-SUBJECT-OUT ACCURACY FOR FULL FEATURE SETS

Feature Set	A_u	A_w
U-FEAT	51.8%	56.5%
B-MFCC	65.9%	67.1%
P-MFCC	63.0%	64.9%
U-FEAT + B-MFCC	72.2%	73.6%
w/Inf. Gain.	**74.6%**	**76.1%**
U-FEAT + P-MFCC	63.6%	65.6%

is expected because many of our features have a high degree of speaker specificity. While this can be mitigated by use of speaker-specific normalization (for example, gender information alone can provide improved results as shown by Tawari [5]), we opted to retain our original feature sets.

We note that in this case, the P-MFCC features were not found to have any advantage over the B-MFCC features. On the other hand, the combination (U-FEAT + B-MFCC) gives almost identical classification results to those reported by Tawari [5]. When feature selection is performed on this configuration to retain features with non-zero information gain, we improve unweighted and weighted accuracy to 74.6% and 76.1%, respectively. This exceeds the figures reported by Tawari [5] for the *leave-one-subject-out* configuration.

VI. CONCLUSION AND FUTURE WORK

Automated emotion recognition from speech requires combining sequences of segmental features with utterance-level features for input to discriminative classifiers. In this paper, we proposed a novel, model-based parametric feature extraction technique for computing feature vectors of high but fixed-dimensionality from a sequence of segmental feature vectors. We argued the potential benefits of this approach, and showed empirically that the proposed features provide significantly better performance than global statistics of segmental features. Using these features, we surpassed the best published 7-way emotion classification accuracy on the *leave-one-text-out* configuration. We also outperformed the state-of-the-art on the *leave-one-subject-out* configuration.

For each feature configuration, we also conducted classification experiments in a reduced 6-dimensional feature space obtained by applying Fisher Multiple Linear Discriminant Analysis (LDA). Model-based features were found to outperform all other feature sets even in this setting. This is a valuable result because it indicates that the increased number of classifier parameters is not the reason for outperformance; rather, the proposed features indeed contain additional information not captured by the averaged segmental feature vectors.

In the present work, we restricted our attention to parametric features obtained from class-conditional GMMs and applied them to the task of emotion recognition. The framework, however, is quite general in that it can be used for many classification problems where segmental features must be combined to make a high-level decision. Further, there is no specific restriction on the underlying probability models. In the future, we plan to experiment with more complex probabilistic time-series models such as HMMs, which may give rise to parametric features of even higher dimensionality. While the current problem dealt with a relatively small number of classes, we also plan to scale our approach to more challenging problems, such as phoneme classification and eventually speech recognition.

ACKNOWLEDGEMENT

This paper is based upon work supported by the DARPA Healing Heroes Program. The views expressed here are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

REFERENCES

[1] A. Ozdas, R. Shiavi, S. Silverman, M. Silverman, and M. Wilkes, "Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 9, pp. 1530–1540, September 2004.

[2] J. Gratch, N. Wang, J. Gerten, E. Fast, and R. Duffy, "Creating rapport with virtual agents," in *Proceedings of the 7th international conference on Intelligent Virtual Agents*, ser. IVA '07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 125–138.

[3] S. Casale, A. Russo, G. Scebba, and S. Serrano, "Speech emotion classification using machine learning algorithms," in *IEEE International Conference on Semantic Computing*, Santa Clara, CA, August 2008, pp. 158–165.

[4] C. Busso, S. Lee, and S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 582–596, May 2009.

[5] A. Tawari and M. Trivedi, "Speech emotion analysis: Exploring the role of context," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 502–509, October 2010.

[6] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "Frame vs. turn-level: Emotion recognition from speech considering static and dynamic processing," in *Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction*, ser. ACII '07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 139–147.

[7] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "Combining frame and turn-level information for robust recognition of emotions within speech," in *Interspeech*, Antwerp, Belgium, August 2007, pp. 2249–2252.

[8] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Interspeech*, Lisbon, Portugal, September 2005, pp. 1517–1520.

[9] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9, pp. 341–345, 2001.

[10] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Networks*, vol. 18, no. 4, pp. 407–422, May 2005.

[11] M. Goudbeek, J. P. Goldman, and K. Scherer, "Emotion dimensions and formant position," in *Interspeech*, Brighton, U.K., 2009, pp. 1575–1578.

[12] A. G. Jr. and D. Wong, "The Burg algorithm for LPC speech analysis/synthesis," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1980.

[13] E. Bozkurt, C. E. Erdem, E. Erzin, T. Erdem, M. Ozkan, and M. Tekalp, "Speech-driven automatic facial expression synthesis," in *3DTV Conference*, Istanbul, Turkey, May 2008, pp. 273–276.

[14] I. T. Nabney, *Netlab: Algorithms for Pattern Recognition*. Springer, 2001.

[15] I. Witten and E. Frank, *Data mining: Practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann, 2005.

[16] J. Platt, *Fast Training of Support Vector Machines using Sequential Minimal Optimization*. MIT Press, January 1998, ch. 12.

[17] C. Bishop, *Neural Networks for Pattern Recognition*. Oxford University Press, 1996.

[18] R. Duin, P. Juszczak, P. Paclik, E. Pekalska, D. de Ridder, D. Tax, and S. Verzakov, *PRTools: A Matlab toolbox for Pattern Recognition*, Delft University of Technology, Delft, Netherlands, 2007.