AUTOMATIC RECOGNITION OF SPEECH EMOTION USING LONG-TERM SPECTRO-TEMPORAL FEATURES

Siqing Wu, Tiago H. Falk, and Wai-Yip Chan

Department of Electrical and Computer Engineering
Queen's University, Kingston, Ontario, Canada
siqing.wu@ece.queensu.ca, {falkt, chan}@ee.queensu.ca

ABSTRACT

This paper proposes a novel feature type for the recognition of emotion from speech. The features are derived from a long-term spectro-temporal representation of speech. They are compared to short-term spectral features as well as popular prosodic features. Experimental results with the Berlin emotional speech database show that the proposed features outperform both types of compared features. An average recognition accuracy of 88.6% is achieved by using a combined proposed & prosodic feature set for classifying 7 discrete emotions. Moreover, the proposed features are evaluated on the VAM corpus to recognize continuous emotion primitives. Estimation performance comparable to human evaluations is furnished.

Index Terms— Emotion recognition, speech processing, spectro-temporal features, affective computing

1. INTRODUCTION

Automatic recognition of human emotions from speech aims at recognizing the underlying emotional state of a speaker from his or her speech signal. It has received rapidly increasing research attention over the past few years, motivated by a broad range of commercially promising applications. While classification of discrete emotions (e.g. *happy*, *neutral*) from speech has been a dominant research focus for some time [1][2][3], recent studies also witness estimation of continuous emotions (e.g. *activation*, *potency*) [4][5].

As a machine learning problem, high performance emotion recognition requires good features. Spectral features play an important role in speech emotion recognition. They are usually extracted over a 20–30 millisecond frame length. Even though longer temporal information can be incorporated in the form of time derivatives, the fundamental character of the features remains quite similar. On the other hand, limitations of short-term features including their derivatives are substantial as described in [6]. Psychoacoustical and neurophysiological findings also indicate the existence of spectro-temporal receptive fields in mammalian auditory cortex which can extend up to temporal spans of hundreds of milliseconds [7], further suggesting the shortcoming of short-term features as they discard the long-term cues used by human listeners. In view of this, we propose a novel feature set in this paper, which is derived from a long-term spectro-temporal (ST) representation of speech.

The speech signal is blocked into long-term windowed segments and a critical-band filterbank is first employed for signal decomposition. Temporal envelopes are then extracted from the decomposed signals of current frame. Lastly, a modulation filterbank is applied to the envelopes to obtain the ST representation of that frame, where modulation frequency is considered jointly with acoustic frequency. In [8], an auditory-inspired ST representation of speech was used to generate a set of preliminary features for speech emotion classification. In this work, we make a more thorough exploration of the ST representation, based on which the proposed features are derived.

Two databases are employed for testing: the Berlin emotional speech database and the Vera am Mittag (VAM) database, which furnish discrete and continuous subjectively assessed emotion descriptors, respectively. The proposed ST features are compared to mel-frequency cepstral coefficient (MFCC) features which are representative short-term spectral features. Popular prosodic features are also extracted to provide performance benchmark. Experimental evaluation indicates that the proposed features are effective for both discrete emotion classification and continuous emotion estimation. They outperform both MFCC and prosodic features, and render a substantial improvement in recognition performance when combined with prosodic features.

2. ST REPRESENTATION OF SPEECH

The auditory spectro-temporal (ST) representation of speech is obtained via the following steps. The input speech signal is first resampled to 8kHz and its active speech level is normalized to -26 dBov using the the P.56 speech voltmeter [9]. Speech frames (without overlap) are labeled as active or inactive by the voice activity detection (VAD) algorithm in [10]; only active speech frames are retained. The preprocessed

speech signal $S(n)$ is framed into segments $S_k(n)$ using a 256 ms Hamming window every 64 ms, where k is the frame index. As described below, the first subband filter in the modulation filterbank performs frequency analysis at frequency contents around 4Hz. Thus this relatively long temporal span is necessary in order to obtain an appropriate frequency resolution. Each speech segment is then processed by two filterbanks. First, a critical-band gammatone filterbank, with N subband filters, is employed to emulate the auditory processing of acoustic signals performed by the human cochlea. The center frequencies of these filters (namely *acoustic* frequency in order to distinguish from *modulation* frequency of the modulation filterbank) are proportional to their bandwidths, which in turn, are characterized by the equivalent rectangular bandwidth [11]. In this work, a filterbank with N=19 filters is used, where the first and the nineteenth filters are centered at 125Hz and 3.5kHz, with bandwidths of 38Hz and 400Hz, respectively.

The Hilbert envelope $\mathcal{H}_k(i,n)$ is then computed from $S_k(i,n)$, which is the output of the ith critical-band filter at frame k ($1 \leq i \leq N$). An M-band modulation filterbank is applied to each $\mathcal{H}_k(i,n)$, generating M outputs $\mathcal{H}_k(i,j,n)$ where j denotes the jth modulation filter ($1 \leq j \leq M$). The filters in the modulation filterbank are second-order bandpass with quality factor set to 2, as suggested in [12]. Here we use an M=5 filterbank whose filter center frequencies are equally spaced on logarithm scale from 4Hz to 64Hz, as it strikes a good balance between performance and model complexity. Lastly, the ST representation of that frame $E_k(i,j)$ is obtained by calculating the energy of $\mathcal{H}_k(i,j,n)$:

$$E_k(i,j) = \sum_{n=1}^{L} \mathcal{H}_k^2(i,j,n), \tag{1}$$

where $1 \leq k \leq T$ with L and T representing the number of samples in one frame and the total number of frames, respectively. For a fixed j, $E_k(i,j)$ relates to the auditory spectral samples of that modulation channel after critical-band grouping. The modulation filterbank allows us to analyze the the modulation frequency content of acoustic frequency components. An example of $E_k(i,j)$ is illustrated in Fig. 1.

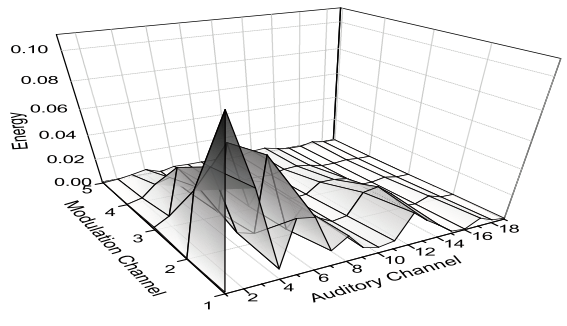

Fig. 1. $E_k(i,j)$ for one frame of a "$neutral$" speech file: low channel index indicates low frequency.

3. FEATURE EXTRACTION

In this section, we describe the proposed features extracted from the ST representation. Short-term spectral features and prosodic features considered in our experiments are also described.

3.1. Spectro-Temporal Features

The ST representation $E_k(i,j)$ is formed via calculating the energy of the Hilbert envelopes. For each frame k, $E_k(i,j)$ is scaled to make it unit energy before further computation, i.e. $\sum_{i,j} E_k(i,j) = 1$. Then four spectral measures Φ_1–Φ_4 are calculated for every modulation channel. For frame k, $\Phi_{1,k}(j)$ is simply the *mean* of the energy samples belonging to the jth modulation channel ($1 \leq j \leq 5$). Φ_1 gives a sense of the energy distribution in speech along the modulation frequency. The second spectral measure is the *spectral flatness* which is conventionally defined as the ratio of the geometric mean of a power spectrum to the arithmetic mean of it. In our calculation, $E_k(i,j)$ is treated as the power spectrum and Φ_2 is thus defined as:

$$\Phi_{2,k}(j) = \frac{\sqrt[N]{\prod_{i=1}^{N} E_k(i,j)}}{\Phi_{1,k}(j)}. \tag{2}$$

A spectral flatness value close to 1 indicates a flat spectrum, while a value close to 0 suggests a spectrum with widely different spectral amplitudes. The third measure employed is the *spectral centroid* which gives a sense of the "center of mass" of the spectrum and is computed as:

$$\Phi_{3,k}(j) = \frac{\sum_{i=1}^{N} f(i) E_k(i,j)}{\sum_{i=1}^{N} E_k(i,j)}. \tag{3}$$

Two types of $f(i)$ have been tried: (1) $f(i)$ being the center frequency (in Hz) of the ith critical-band filter of the auditory filterbank, and (2) $f(i) = i$. No remarkable difference in performance can be observed; the latter is chosen as it has a simpler form. Moreover, given the observation that adjacent modulation channels usually have considerable correlation, spectral flatness and centroid of adjacent modulation channels also exhibit high correlation. In order to alleviate such information redundancy, Φ_2 and Φ_3 are computed for $j \in \{1,3,5\}$ only. Additionally, we group the 19 acoustical bins into four divisions: 1—4, 5—10, 11—15, 16—19, namely D_l ($1 \leq l \leq 4$). Bins within each division are grouped together: $\mathbb{E}_k(l,j) = \sum_{i \in D_l} E_k(i,j)$. Then the *modulation spectral centroid* (Φ_4) is calculated in a way essentially the same as Φ_3:

$$\Phi_{4,k}(l) = \frac{\sum_{j=1}^{M} j \mathbb{E}_k(l,j)}{\sum_{j=1}^{M} \mathbb{E}_k(l,j)}. \tag{4}$$

By calculating Φ_4, we extract the information about how fast the selected acoustic frequency region varies, in other words,

the temporal dynamic cues. In total, 15 features are obtained from the ST representation per frame by means of the four spectral measures.

Furthermore, linear predication (LP) analysis is applied to selected modulation channels where $j \in \{1, 3, 5\}$, to extracted another set of ST features from $E_k(i, j)$. The autocorrelation method is used. In order to suppress local details while preserving the broad structure beneficial to recognition, a 5th-order all-pole model is empirically used to approximate the spectral samples. The computational cost of this autoregressive modeling is negligible due to the low LP order and the limited number of spectral samples per modulation channel (19 here). The LP coefficients are further transformed into cepstral coefficients (LPCC) and denoted as $C_k(n, j)$ ($0 \leq n \leq 5$). The LPCCs have shown to be a generally more robust, reliable feature set for speech recognition than the direct LP coefficients. Together with the 15 aforementioned features, in total 33 ST features are calculated frame-by-frame.

While raw speech features are extracted on frame-level (FL), the most common practice among current works in emotion recognition is to apply functions (usually statistical) to the time trajectories/contours of the FL features, to derive utterance-level (UL) features [2–5, 13–17]. In [16], the UL method is shown to ourperform FL dynamic modeling for emotion recognition. The superiority of the UL method mainly comes from the fact that it avoids spoken-content over-modeling [17]. Consequently, mean and standard deviation (std. dev.) of the FL ST features are calculated, giving 66 UL ST features. They are denoted as S_{66}.

3.2. Short-Term Spectral Features

The mel-frequency cepstral coefficients (MFCCs) are extensively used short-term spectral features in speech recognition. They are extracted in this work to compare to the proposed long-term features. The preprocessed speech signal is first filtered by a high-pass filter with pre-emphasis coefficient 0.97, and the first 13 MFCCs (including the zeroth order coefficient) are extracted from 25 ms Hamming-windowed speech frames every 10 ms. The most commonly used MFCC features amongst current works are mean and std. dev. (or variance) of the first 13 MFCCs and their deltas [4][5][13] [14][17]. In this work, we compute mean, std. dev., and the 3rd up to the 5th central moments of the 13 MFCCs and their deltas and double-deltas, resulting in 195 MFCC features. This set is akin to the one used for comparison in [13], except that it further considers the delta coefficients. Denote it as M_{195}.

3.3. Prosodic Features

Prosodic features have been, among numerous acoustic features used for speech emotion recognition, a standard feature type in previous works. However, the "best" set of prosodic features has yet to be found and may in reality depend on specific application. Consequently, the prosodic features used by recent works differ considerably [14][15]. Nevertheless, the state-of-art way of deriving prosodic features relies on applying functions to trajectories/contours of pitch, energy, and sometimes also their deltas, etc. In this vein, the trajectories of pitch and intensity (in dB), and their deltas are extracted in this work. Then the following statistics are computed for each trajectory: mean, std. dev., skewness, kurtosis, maximum, minimum, quartiles, range, and differences between quartiles. The linear and quadratic regression coefficients of the trajectories plus the root mean squared error, are further calculated as features. Moreover, mean and std. dev. of syllables' durations, and ratio between voiced and unvoiced segments are also measured. In total, 71 prosodic features are extracted. Denote them as P_{71}.

4. DATA

4.1. Berlin Emotional Speech Database

The Berlin emotional speech database [18] is used for experiments classifying discrete emotions. Ten actors (5m/5f) each uttered ten sentences (5 short and 5 longer, typically between 1.5 and 4 seconds) in German to simulate 7 different emotions. Utterances scoring higher than 80% emotion recognition rate in a subjective listening test are included in the database. We classify all the 7 emotions in this work. The numbers of speech files for these emotion categories in the presented Berlin database are: *anger* (127), *boredom* (81), *disgust* (46), *fear* (69), *joy* (71), *neutral* (79) and *sadness* (62).

How the specific language (or database) affects the recognition performance remains an open issue in emotion recognition, and is beyond the scope of this paper. The Berlin database is employed here because it is one of the most popular databases used by researchers on emotion recognition, thereby facilitating the comparison with other works. However, as suggested by the preliminary results in [3], the trained emotion classifiers are usually highly language-dependent, therefore, appropriate language adaptation has to be performed if the application involves multiple languages.

4.2. Vera am Mittag Database

The VAM database [19] is a speech corpus of spontaneous emotions. It was recorded from a German TV talk-show. The recordings are manually segmented at the utterance level. The presented VAM database contains two parts: VAM I of 478 utterances from 19 speakers (4m/15f) with 17 human evaluators, and VAM II of 469 utterances from 28 speakers (7m/21f) with 6 evaluators. Three emotion primitives: *valence*, *activation* and *dominance*, are assessed by the evaluators. Primitive values are normalized to the range of $[-1, +1]$. Correlation between the evaluators and estimation error are calculated for each primitive as described in [4].

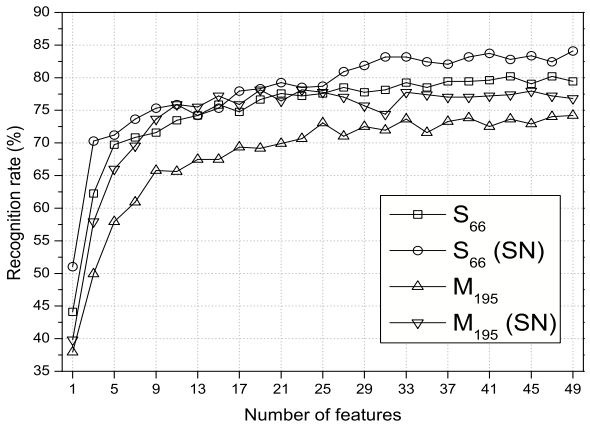

Fig. 2. Comparison between ST and MFCC features.

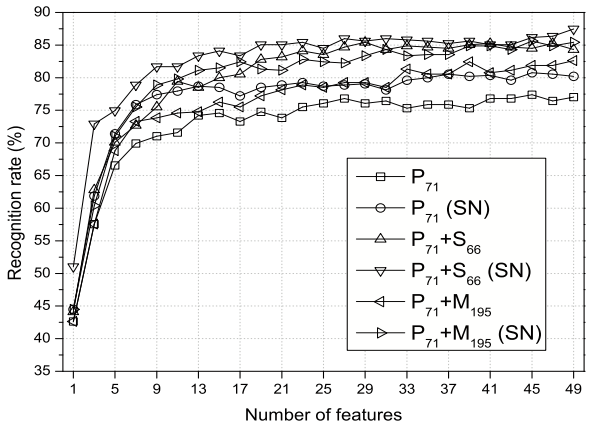

Fig. 3. Comparison between prosodic and combined features.

5. RESULTS

In this section, results of experimental evaluation are presented. The proposed ST features are first compared to MFCC features. Then their contribution as supplementary features to prosodic features is studied. The support vector machines with radial basis function (RBF) kernels are used for discrete emotion classification as well as continuous emotion primitive estimation. The implementation in [20] is adopted. The features from training data are linearly scaled to $[-1, 1]$ before applying SVM, with features from test data scaled using the trained linear mapping function.

5.1. Comparison with MFCC Features

All results achieved on the Berlin database are produced using 10-fold cross-validation. The proposed ST features are first compared to MFCC features (S_{66} vs. M_{195}). The effect of taking a speaker normalization (SN) step before recognition is also investigated. For SN, the features are processed by the mean and variance normalization (MVN) within the scope of each speaker [17] to compensate speaker variations prior to applying SVM. The well-known sequential forward feature selection (SFS) [21] is used to select the most salient features. It finds subsets of the original features. Classification results are shown in Figure 2 where a varying number of features (up to 50) are selected by SFS[1]. It is clear from the figure that without SN, the proposed features consistently outperform MFCC features by a wide margin. But MFCC features benefit more from SN, especially when a small number of features are selected. Nevertheless, the proposed features still prevail as more features are included and offer the best result (84.1% for S_{66} vs. 78.7% for M_{195}, both with SN).

[1]The accuracy trajectories are downsampled by a factor of 2 for visual purpose, i.e. only values at odd feature number are shown.

5.2. Comparison with Prosodic Features

As aforementioned, prosodic features are probably the most widely used features in current works on speech emotion recognition, and thus offer a baseline performance. Therefore, it is also of our interest to study the contribution of the proposed features in addition to prosodic features. We conduct tests using: (1) only prosodic features, and (2) combined prosodic & ST features. MFCC features are still considered for comparison purpose. Classification results are shown in Figure 3. As indicated by Figures 2 and 3, when each feature type is tested individually, prosodic features generally give better results than MFCC features, and the proposed features outperform prosodic features. Adding spectral features to prosodic features is shown to be useful as better performance is achieved. The ST features still outperform MFCC features after feature combination, regardless of SN. But MFCC features may be more complementary to prosodic features than the proposed features, as we can see that the advantage of the ST features over MFCC features decreases after feature combination (cf. Fig. 2). The combined prosodic & ST features with SN offer the best performance, and up to 88.6% recognition accuracy can be achieved. A list of top features selected from P_{71}+S_{66} (SN) by SFS is available in the Appendix.

Tables 1 and 2 show the confusion matrices (left-most column being the true emotions) for the best recognition performance achieved by prosodic features only (45 features, SN, 80.8%) and combined prosodic & ST features (50 features, SN, 88.6%), respectively. We can see from the tables that adding ST features contributes to the recognition of all emotion types. Most emotions can be recognized well except *joy*. The most notable confusion pair is shown to be *joy* and *anger*, although they are of opposite valence. This might be due to the fact that *activation* is more easily recognized by machine than *valence*, as indicated by the regression results for the emotion primitives on the VAM database in the next section.

In [13][2], multi-stage classification technique is exper-

[2]Unless otherwise specified, all following cited results are achieved on

Table 1. Recognition rates with *prosodic* features.

Emotion	Anger	Boredom	Disgust	Fear	Joy	Neutral	Sadness	Rate
Anger	**109**	0	5	0	12	1	0	**85.8%**
Boredom	1	**68**	2	0	0	3	7	**84.0%**
Disgust	4	1	**37**	1	1	2	0	**80.4%**
Fear	8	0	1	**53**	5	2	0	**76.8%**
Joy	26	0	2	4	**37**	2	0	**52.1%**
Neutral	1	1	3	0	0	**70**	4	**88.6%**
Sadness	0	2	0	0	0	2	**58**	**93.6%**
Precision	**73.2%**	**94.4%**	**74.0%**	**91.4%**	**67.3%**	**85.4%**	**84.1%**	

Table 2. Recognition rates with *combined* features.

Emotion	Anger	Boredom	Disgust	Fear	Joy	Neutral	Sadness	Rate
Anger	**120**	0	1	1	5	0	0	**94.5%**
Boredom	0	**73**	0	0	1	4	3	**90.1%**
Disgust	0	2	**39**	1	1	3	0	**84.8%**
Fear	2	0	0	**61**	6	0	0	**88.4%**
Joy	19	0	2	2	**47**	1	0	**66.2%**
Neutral	1	2	1	1	0	**73**	1	**92.4%**
Sadness	0	1	0	0	0	0	**61**	**98.4%**
Precision	**84.5%**	**93.6%**	**90.7%**	**92.4%**	**78.3%**	**90.1%**	**93.9%**	

imented and shown to be useful, as one-stage, two-stage and three-stage classification schemes give 75%, 83.5%, and 88.8% accuracy for classifying 6 emotions (no *disgust*), respectively. In [17], 83.2% recognition rate is reported using roughly 1.4k UL candidate features, however, under a speaker-independent condition which is more stringent, and by further integrating frame-level information, the accuracy is improved to 89.9%. In [22], a novel type of long-term modulation features has also been experimented, which achieves roughly 70% recognition rate when combined with other two feature types.

5.3. Results on VAM Database

The proposed ST features are further examined on the VAM database to estimate continuous emotion primitives. Leave-one-out (LOO) cross-validation is used to facilitate comparison with the results in [4]. Regression results using prosodic, proposed and combined features (without SN) are shown in Table 3, where r and e stand for correlation and mean absolute error, respectively. The features selected on the Berlin database are used here[3]. The machine recognition and human subjective evaluation results reported in [4] are also included for comparison[4].

As shown in Table 3, the proposed features give higher correlations for estimating *activation* and *dominance* than prosodic features. But both features give poor estimation of *valence*. Adding ST features is still useful as combined features give the highest correlations on all the three datasets.

the Berlin database for classifying 7 emotions.

[3]Since training data differs from trial to trial for cross-validation, features ranked within top 50 by SFS more than twice during the 10 trails are used.

[4]In [4], only standard deviation of subjective evaluation is presented. The error values listed here are inferred values.

The proposed algorithm yields a smaller estimation error compared to the machine estimation results in [4]. The performance is even somewhat better than human assessment. Amongst the three primitives, *activation* is shown to be most easily estimated with up to 0.86 correlation achieved on VAM I, followed by *dominance*, and *valence* shows significantly inferior correlation even though its estimation error is small. This verifies our previous assumption that *activation* can be more easily recognized by machine than *valence*, and also is in accordance with human evaluation results. In [5], good estimation results are also achieved for *activation* and *dominance* on VAM I+II, but *valence* is still poorly estimated (0.46 correlation).

6. CONCLUSION

In this paper we propose a novel feature set for speech emotion recognition. The features are derived from a long-term ST representation of speech. They are shown to outperform both MFCC and conventional prosodic features, and can serve as useful additions to prosodic features.

7. ACKNOWLEDGEMENT

The authors would like to thank Dr. Michael Grimm for his support and making the VAM database available.

8. APPENDIX: LIST OF FEATURES

Denote the feature set as F. The average feature rank (AFR) of feature $f_i \in F$ given R trails is calculated as:

$$\mathrm{AFR}(f_i) = \frac{1}{R}\sum_{r=1}^{R} \text{rank of } f_i \text{ in the } r\text{th trial}. \qquad (5)$$

If f_i is not selected, its rank is replaced by a penalty value P. The top 10 features from the combined ST & prosodic feature set are listed in Table 4 as ranked by AFR, with P set to 11.

Table 4. Top 10 combined features ranked by AFR.

Rank	Feature	AFR	Rank	Feature	AFR
1	mean of $\Phi_{3,k}(1)$	1.0	6	mean syllable duration	8.3
2	mean of $\Phi_{1,k}(3)$	2.0	7	slope of intensity	8.7
3	mean of pitch	3.8	8	mean of $\Phi_{4,k}(2)$	9.6
4	std. dev. of pitch	6.9	9	mean of $C_k(3,1)$	9.7
5	mean of $\Phi_{4,k}(3)$	7.0	10	range of intensity	9.8

9. REFERENCES

[1] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: towards a new generation of databases," *Speech Communication*, vol. 40, pp. 33–60, 2003.

Table 3. Correlation and absolute error for emotion primitive regression on the VAM database.

Dataset	Feature Type	Correlation (r)			Absolute Error (e)			Average	
		valence	activation	dominance	valence	activation	dominance	$\overline{r}$	$\overline{e}$
VAM I	Prosodic	0.50	0.81	0.77	0.12	0.17	0.16	0.69	0.15
	Proposed	0.56	0.84	0.78	0.11	0.16	0.15	0.73	0.14
	Combined	0.58	0.86	0.81	0.11	0.15	0.15	0.75	0.14
	Results in [4]	N/A	N/A	N/A	N/A	N/A	N/A	0.71	0.27
	Human	0.49	0.78	0.68	0.17	0.25	0.20	0.65	0.21
VAM II	Prosodic	0.29	0.62	0.57	0.15	0.18	0.16	0.49	0.16
	Proposed	0.24	0.71	0.60	0.15	0.17	0.16	0.52	0.16
	Combined	0.32	0.72	0.63	0.15	0.16	0.16	0.56	0.16
	Results in [4]	N/A	N/A	N/A	N/A	N/A	N/A	0.43	0.23
	Human	0.48	0.66	0.54	0.11	0.19	0.14	0.56	0.15
VAM I+II	Prosodic	0.43	0.73	0.70	0.13	0.18	0.16	0.62	0.16
	Proposed	0.42	0.80	0.73	0.13	0.16	0.16	0.65	0.15
	Combined	0.46	0.81	0.77	0.13	0.16	0.15	0.68	0.15
	Results in [4]	N/A	N/A	N/A	N/A	N/A	N/A	0.60	0.24
	Human	N/A	N/A	N/A	N/A	N/A	N/A	0.61	0.18

[2] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, pp. 1162–1181, 2006.

[3] M. Shami and W. Verhelst, "An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech," *Speech Communication*, vol. 49, pp. 201–212, 2007.

[4] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, pp. 787–800, 2007.

[5] M. Grimm, K. Kroschel, and S. Narayanan, "Support vector regression for automatic recognition of spontaneous emotions in speech," in *ICASSP*, 2007, vol. 4, pp. 1085–1088.

[6] N. Morgan *et al*., "Pushing the envelope - aside," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 81–88, 2005.

[7] T. Chih, P. Ru, and S. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *J. Acoust. Soc. Amer.*, vol. 118, pp. 887–906, 2005.

[8] S. Wu, T. H. Falk, and W.-Y. Chan, "Long-term spectrotemporal information for improved automatic speech emotion classification," in *Interspeech*, Brisbane, Australia, 2008.

[9] ITU-T P.56, "Objective measurement of active speech level," Intl. Telecom. Union, Switzerland, 1993.

[10] ITU-T G.729 Annex B, "A silence compression scheme for g.729 optimized for terminals conforming to itu-t recommendation v.70," Intl. Telecom. Union, Switzerland.

[11] B. Glasberg and B. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Res.*, vol. 47, pp. 103–138, 1990.

[12] S. D. Ewert and T. Dau, "Characterizing frequency selectivity for envelope fluctuations," *J. Acoust. Soc. Amer.*, vol. 108, pp. 1181–1196, 2000.

[13] M. Lugger and B. Yang, "Cascaded emotion classification via psychological emotion dimensions using a large set of voice quality parameters," in *ICASSP*, 2008, vol. 4, pp. 4945–4948.

[14] B. Schuller *et al*., "The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals," in *Interspeech*, 2007, pp. 2253–2256.

[15] A. Batliner *et al*., "Combining efforts for improving automatic classification of emotional user states," in *Fifth Slovenian and First International Language Technologies Conference (IS-LTC)*, 2006, pp. 240–245.

[16] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in *ICASSP*, 2003, vol. 2, pp. 1–4.

[17] B. Vlasenko *et al*., "Combining frame and turn-level information for robust recognition of emotions within speech," in *Interspeech*, 2007, pp. 2225–2228.

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

[19] M. Grimm, K. Kroschel, and S. Narayanan, "The vera am mittag german audio-visual emotional speech database," in *ICME*, 2008, pp. 865–868.

[20] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," Tech. Rep., 2001, Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[21] J. Kittler, "Feature set search algorithms," *Pattern Recognition and Signal Processing*, pp. 41–60, 1978.

[22] S. Scherer, F. Schwenker, and G. Palm, "Classifier fusion for emotion recognition from speech," in *3rd IET International Conference on Intelligent Environments (IE 07)*, 2007, pp. 152–155.