EMOTION CLASSIFICATION OF SPEECH USING MODULATION FEATURES

Theodora Chaspari [1], Dimitrios Dimitriadis [2], Petros Maragos [3]

[1]USC EE Dept, Los Angeles, CA, USA, [2]AT&T Labs Research, Florham Park, NJ, USA
[3] NTUA School of ECE, Athens, Greece

chaspari@usc.edu, ddim@research.att.edu, maragos@cs.ntua.gr

ABSTRACT

Automatic classification of a speaker's affective state is one of the major challenges in signal processing community, since it can improve Human-Computer interaction and give insights into the nature of emotions from psychology perspective. The amplitude and frequency control of sound production influences strongly the affective voice content. In this paper, we take advantage of the inherent speech modulations and propose the use of instant amplitude- and frequency-derived features for efficient emotion recognition. Our results indicate that these features can further increase the performance of the widely-used spectral-prosodic information, achieving improvements on two emotional databases, the Berlin Database of Emotional Speech and the recently collected Athens Emotional States Inventory.

Index Terms— Emotion classification, AM-FM features, speech analysis, human-computer interaction

1. INTRODUCTION

Automatic emotion recognition has recently gained a lot of interest in the signal processing field. It focuses on the development of techniques to automatically recognize human emotional states aiming at making the human-computer interaction more natural and giving further insights into emotional expression [1].

Research efforts in this field have focused on both feature extraction and classification. Low-level contours of prosody, voice quality and articulatory information have been used to extract high-level functionals that describe the emotional content of a sentence [2]. Diverse time scales of frame- and turn-level have been also combined to recognize emotions from speech [3]. Subsequent layers of binary classifiers were further fused in a hierarchical framework for emotion classification [4] and emotion profiles were introduced in order to identify emotional properties of ambiguous utterances [5]. In the context of continuous tracking of affective states, a Gaussian Mixture Model-based approach [6] and a long short-term memory neural network [7, 8] have also been proposed.

Although speech emotion recognition has mainly focused on the extraction of prosodic and spectral features [3,9], other studies have attempted to use the modulation properties of speech signals [10, 11] to recognize human affective states. Our approach differs from [11], in that we use multiple frequency bands to compute amplitude and frequency modulations based on the Energy Separation Algorithm (ESA) [12]. We also extend the ideas proposed in [10] for stress recognition including multiple AM-FM metrics, such as the instantaneous amplitude and frequency mean and standard deviation.

In this paper we introduce the use of short term speech modulations [12], referred here as "Micro-Modulations". A speech signal can be modeled as a sum of AM-FM resonances of the type:

$$r_i(t) = \alpha_i(t) \cos\left(2\pi \int_0^t f_i(\tau) d\tau\right)$$

each with instantaneous amplitude $\alpha(t)$ and instantaneous frequency $f(t)$ [12]. To compute these AM-FM signals the Gabor ESA [13] is used. Similar modulation-inspired features were shown to benefit speech recognition as well [14].

Our proposed modulation features are compared with prosodic-spectral descriptors commonly used in these tasks within different classification frameworks. We validate our approach with two acted emotional databases: Berlin Database of Emotional Speech (EmoDB) [15] and Athens Emotional States Inventory (AESI). The later is a new database created under the collaboration of National Technical University and Mental Health Care Unit of Athens for the purposes of this paper. Our experiments show that the combination of AM-FM and spectral features can provide accuracies up to 79.8% and 67.4% for the two considered databases respectively giving relative improvement of 10.7% and 12% upon the baseline. Although better results have been reported in EmoDB [3, 9, 16], the much lower dimensionality of AM-FM features compared to the ones described in literature [3, 16] and the fact that their evaluation was performed with a more strict experimental setup [9, 16] suggest that the proposed features consist a viable front-end framework for emotion recognition.

2. MODULATION FEATURES FOR EMOTION

In this section the proposed feature extraction scheme is detailed (Fig. 1). Let $x[k]$ be the original speech signal and $x_i[k]$ be the signal corresponding to the i^{th} frequency band ($k = 1, \ldots, K$), where K is the frame length in samples.

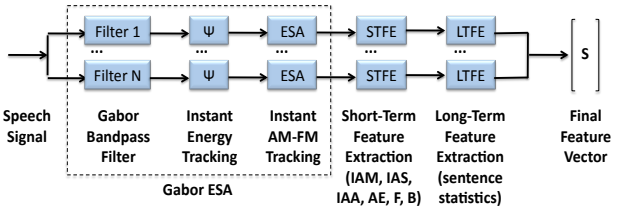

Fig. 1. *Feature extraction overview.*

We use ESA to compute its instantaneous amplitude and frequency signals as follows: $|\alpha_i[k]| \approx \frac{\Psi(x_i[k])}{\sqrt{\Psi(\dot{x}_i[k])}}$ and $f_i[k] \approx \frac{1}{2\pi}\sqrt{\frac{\Psi(\dot{x}_i[k])}{\Psi(x_i[k])}}$, where $\Psi(x) = \dot{x}^2 - x\ddot{x}$ is the Teager-Energy operator. Ψ is modeled continuously but implemented via discrete convolutions with derivatives of the i^{th} Gabor bandpass filter's impulse response g_i; this increases robustness of the ESA [13]. Alternatively, a Gammatone filterbank can be used [17]. Thus,

$$\Psi(x_i[k]) = (x[k] * \dot{g}_i[k])^2 - (x[k] * g_i[k])\,(x[k] * \ddot{g}_i[k])$$

$$\Psi(\dot{x}_i[k]) = (x[k] * \ddot{g}_i[k])^2 - (x[k] * \dot{g}_i[k])\,(x[k] * \dddot{g}_i[k])$$

Due to space limitation, in the above equations we use the compact notation $x[k] * \dot{g}[k] = x[k] * \frac{dg(t)}{dt}|_{t=kT}$, where T is the sampling period, applying also to higher order derivatives.

To suppress the possible fluctuations, we performed 7-sample median filtering and derived short-term features over a frame of 25ms length (corresponding to K samples, as mentioned previously) with 10ms overlap. All features were computed over 4-24 (with step 4) Mel-scale distributed bands.

"Micro-Amplitude" and "Micro-Frequency" features [14] were derived from instant amplitude and frequency signals computed with ESA over each frame $m = 0, \ldots, M-1$, where M is the number of speech frames. The term "micro" refers to their inherent short-term information. They include instant amplitude mean and standard deviation defined as:

$$IAM_i[m] = \tfrac{1}{K}\textstyle\sum_{k=mK+1}^{(m+1)K} |\alpha_i[k]|$$

$$IAS_i[m] = \sqrt{\tfrac{1}{K-1}\textstyle\sum_{k=mK+1}^{(m+1)K} (|\alpha_i[k]| - IAM_i[m])^2}$$

Inspired by the features introduced in [10], we computed the area under the instant envelope $e_i(t)$, which is a median smoothed version of $\alpha_i(t)$, and its autocorrelation area:

$$IAA_i[m] = \textstyle\sum_{k=mK+1}^{(m+1)K} e_i[k]$$

$$AE_i[m] = \textstyle\sum_{k=mK+1}^{(m+1)K} e_i[k] * e_i[-k]$$

We also computed the weighted mean F and square bandwidth B^2 of instant frequency defined as:

$$F_i[m] = \frac{\sum_{k=mK+1}^{(m+1)K} f_i[k]\cdot(\alpha_i[k])^2}{\sum_{\ell=mK+1}^{(m+1)K} (\alpha_i[\ell])^2}$$

$$B_i^2[m] = \frac{\sum_{k=mK+1}^{(m+1)K} \left[(\dot{\alpha}_i[k])^2 + (f_i[k]-F_i[m])^2(\alpha_i[k])^2\right]}{\sum_{\ell=mK+1}^{(m+1)K} (\alpha_i[\ell])^2}$$

To represent the time evolution of AM-FM signals, we used the first derivative of IAM, IAA, AE and F, which will be referred as IAM-Der, IAA-Der, AE-Der and F-Der respectively. Second-order time derivatives were proven to be too noisy and were not included.

Although there are many possible AM-FM feature combinations, we indicatively used the following:

- Mod1: IAM + IAM-Der + IAS + F + F-Der + B^2
- Mod2: IAA + IAA-Der + IAS + F + F-Der + B^2
- Mod3: AE + AE-Der + IAS + F + F-Der + B^2

In an attempt to capture global trends, we computed the mean and standard deviation of the above features over each emotional sentence. This results in 12 features per band for each of the 4-24 frequency bands.

The described features contain non-linear acoustic information due to the modulations. We combine these with the mean and standard deviation, computed over each sentence, of the first 13 MFCCs and their first-order derivatives (noted as Mod1+MFCC, Mod2+MFCC, Mod3+MFCC), in order to capture the linear source-filter acoustic information. This results in a feature array of 100-340 dimensions depending on the number of frequency bands.

There is evidence from auditory physiology, psychoacoustics and speech perception, that signal modulations are related to sound perception [18]. Concerning the analysis of speech emotional content, it was found that amplitude modulation substantially enhances the information transmitted in human speech [19]. We indicatively examined the influence of emotion to instant amplitude and for the sake of simplicity we selected four emotions that are referred as bipolar pairs according to Plutchik's wheel representation [20]: anger and fear, joy and sadness. Low activated emotions, such as fear and sadness, tend to have stronger amplitude modulations in low frequencies, whereas anger and joy depict strong amplitude content through the whole frequency spectrum, as demonstrated in our data. The instant amplitude mean over time and modulation frequencies for an given sentence in EmoDB uttered from the same speaker is shown in Fig. 2.

3. FEATURE SELECTION AND CLASSIFICATION

Since some of the modulation features might have correlated information, we transformed the original proposed and baseline feature sets to improve classification accuracy.

Speaker normalization was used to compensate for speaker variability by normalizing the features from same-speaker sentences to have zero mean and unity standard deviation.

To increase feature discriminability, we performed feature selection according to the Fisher Discriminant Ratio criterion [21], defined as:

$$FDR = \tfrac{1}{N(N-1)}\textstyle\sum_{n=1}^{N}\sum_{\ell=n+1}^{N} \frac{(\mu_n-\mu_\ell)^2}{\sigma_n^2-\sigma_\ell^2}$$

where N is the number of emotional classes and μ_n, σ_n^2 are class n mean and variance. If a feature's FDR value was less

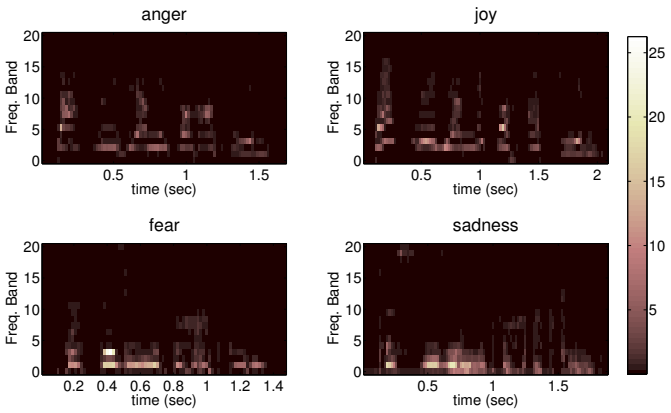

Fig. 2. *Visualization of instant amplitude mean (as computed for 20 frequency bands) over time and frequency for sentence a02 of EmoDB ("Das will sie am Mittwoch abgeben", i.e. "She will hand it in on Wednesday") uttered by speaker 13.*

than a percentile threshold f_{prc} of the maximum FDR of the original feature set, the corresponding feature was omitted.

A decorrelation scheme was used to avoid redundant information, mainly because of the overlap in the successive frequency bands. If the Pearson's correlation coefficient for a given pair of features is higher than a threshold value r_{thr}, we included only the feature with higher FDR.

Finally, we performed LDA over the feature set from decorrelation to further increase class separation. Since we have 7 and 5 classes in EmoDB and AESI respectively and LDA results in one fewer dimension than the total classes [21], we get a final 6- and 4-dimensional feature vector per dataset.

The same selected features were fed to two classifiers, i.e. Gaussian Mixture Models (GMMs) and Support Vector Machines (SVMs). GMMs had a full covariance matrix and were implemented with HTK [22]. We used binary SVMs with linear kernel trained with SVM-light [23] for each pair of classes and took the final decision based on majority voting for the class chosen by the maximal number of SVMs.

4. EXPERIMENTS

4.1. Data Description

Since the effect of emotion on the AM-FM features has not been extensively studied, we used two databases with emotional sentences of acted speech in order to keep our experiments as controlled as possible.

EmoDB [15] includes 535 utterances from seven basic emotions (anger, boredom, disgust, fear, happiness, sadness, neutral). Ten native German professional actors expressed ten different sentences each with all emotions. 65 sentences were omitted due to annotator disagreement.

AESI contains 696 utterances of five basic emotions (anger, joy, sadness, fear, neutral) from 20 native Greek students. The content of each sentence reflects the corresponding emotion and was validated based on a systematic survey with 40 people (separate from the ones participating in the recording) who were asked to rate each sentence based on the 5 predetermined emotional states. The recordings were performed in a sound-proof room using the same setup and audiovisual equipment with 44.1kHz sampling rate. Each student expressed 7 sentences per emotion, resulting in balanced number of samples across classes (~ 140 sentences/emotion). In total, 4 utterances were omitted by the speakers by accident. The originality of the database occurs in the fact that it is of the few systematic efforts to record emotional sentences in greek with high quality audiovisual equipment. Previous studies focused on recording isolated greek emotional words [24] or more unstructured data [25].

4.2. Experimental details

Our classification scheme included a leave-one-speaker-out cross-validation. In each fold one speaker was used for test, one for development (dev) and the rest for train set. Dev-set was used for tuning feature selection and classifier parameters to avoid overfitting. The values tested were $f_{prc} = 0.05, \ldots, 0.5$ (with step 0.05) for the FDR percentile and $r_{thr} = 0.1, \ldots, 0.9$ (with step 0.1) for the Pearson correlation threshold. The system had to choose between 1 or 2 Gaussian mixtures for GMM and among $C \in \{0.01, 0.1, 1, 5, 10\}$ trade-off parameter values for SVM. All feature transformations were derived based on the train data and then applied on the dev and test set for each fold.

The baseline against which we compared our proposed modulation features consists of the first 13 Mel-Frequency Cepstral Coefficients (MFCCs), their first-order derivatives and the fundamental frequency of speech, resulting in 27 features. We computed their mean and standard deviation over the whole sentence, consisting an array of 54 parameters. We will refer to this as "MFCC+Prosody". Despite its simplicity, this baseline is able to capture part of the spectral and prosodic information relevant for our task. Since it is much lower in dimensionality than other commonly used frameworks, such as descriptors extracted with the openSMILE toolkit [26], it is more comparable to our AM-FM approach. In future work we plan to assess the performance of AM-FM features with other more complicated front-end systems.

The proposed modulation feature sets (Mod1, Mod2, Mod3), their combination with MFCCs (Mod1+MFCC, Mod2+MFCC, Mod3+MFCC) and the baseline (MFCC+ Prosody) were evaluated with the two classifiers.

4.3. Results

Our system performance is measured with the unweighted classification accuracy (UA), a less sensitive measure to unbalanced distribution of instances among classes defined as:

$$UA = \frac{1}{N} \sum_{n=1}^{N} \frac{c_n}{\nu_n}$$

where N is the number of emotional classes, ν_n is the number of samples in class n and c_n is the number of correctly classified samples from class n.

Table 1. *Unweighted classification accuracy (UA, %) in EmoDB and AESI using two classifiers (GMM, SVM) based on the three groups of modulation features (Mod1, Mod2, Mod3) computed over 4-24 (with step 4) frequency bands, their combinations with MFCCs (Mod1+MFCC, Mod2+MFCC, Mod3+MFCC) and the baseline (MFCC+Prosody). Significant difference between the median UA over all folds of the proposed against the baseline system was evaluated with a one-sided Wilcoxon Rank-Sum test (*,†: statistical significance $p < 0.05, 0.1$, respectively). Bold font indicates the best result for each feature group and classifier.*

Database	Feature	GMM						SVM					
		Number of Freq. Bands						Number of Freq. Bands					
		4	8	12	16	20	24	4	8	12	16	20	24
EmoDB	**MFCC+Prosody**	71.3						72.9					
	Mod1	51.6	55.8	64.4	58.3	**66.4**	63.1	48.1	56.1	62.9	60.6	59.7	58.5
	Mod2	51.8	55.8	64.4	58.2	62.1	63.4	46.6	54.0	60.9	57.5	63.0	57.8
	Mod3	54.8	53.9	62.1	61.9	61.6	64.8	50.7	54.1	59.7	59.3	**66.3**	60.4
	Mod1+MFCC	77.7	74.9	76.2	79.4†	79.2*	77.6†	72.0	73.7	74.8	74.6	**77.0**	74.7
	Mod2+MFCC	77.7	74.9	76.3	79.4†	79.2*	77.3†	71.2	74.3	73.7	74.0	76.4	74.6
	Mod3+MFCC	74.6	75.7	78.0†	77.8†	**79.8***	77.8†	73.7	69.4	73.6	73.6	76.2	74.3
AESI	**MFCC+Prosody**	60.2						49.2					
	Mod1	53.2	50.8	**59.9**	57.1	52.4	54.7	42.0	39.2	43.0	44.8	44.3	44.7
	Mod2	52.4	50.7	59.8	56.9	52.4	55.4	37.5	43.4	38.1	45.7	43.8	**45.9**
	Mod3	53.4	52.4	54.3	56.3	50.3	54.7	40.6	40.7	42.0	40.1	43.8	44.3
	Mod1+MFCC	63.7	65.0†	65.7*	67.0*	64.8†	64.6	55.4*	57.3*	58.8*	**59.3***	58.9*	57.2*
	Mod2+MFCC	63.7	65.1†	65.7*	**67.4***	65.0†	64.6	57.7*	58.1*	53.9†	57.8*	53.8†	57.0*
	Mod3+MFCC	64.9*	65.3†	66.1*	67.2*	65.3†	63.4	58.5*	56.6*	57.3*	55.1*	56.6*	53.2

UAs of modulation features across all frequency bands, their combination with MFCCs and the MFCC+Prosody baseline are reported in Table 1. Significance of improvement is evaluated with a one-sided Wilcoxon Rank-Sum test [27] comparing the medians of classification accuracies over all cross-validation folds between proposed and baseline system.

Increasing the number of frequency bands yields in better accuracies, since we incorporate more detailed information of speech modulations. AM-FM features computed over many frequency bands provide lower results but comparable to the spectral-prosodic ones and improve performance upon baseline when combined with MFCCs. UAs reach up to 79.8% (Mod3+MFCC, 20 freq. bands, 292 features before selection) and 67.4% (Mod2+MFCC, 16 freq. bands, 244 features) in EmoDB and AESI respectively. Performance on EmoDB is higher, which could stem from AESI's variable lexical content. AESI utterances were different for each emotion, which was not the case for EmoDB. Mod2 and Mod3 containing instant amplitude area and Autocorrelation Envelope (both derived after instant amplitude smoothing), tend to yield better results. Finally, GMMs give higher UAs than SVMs, which could be due to GMMs' ability to better model the low dimensional feature space resulting from LDA.

Although better performance has been reported in EmoDB, the corresponding studies use more complex front-end frameworks or employ less strict experimental setup. Schuller et al. [28] achieve UAs up to 85.6% with 6552 features of prosodic and spectral supra-segmental information. Using a similar scheme (5967 features) and splitting two groups of speakers for the train and test set, unweighted precision and recall reach 76.1% and 83.6% in a 6-class classification task (omitting disgust) [29]. Cepstral low-level attributes integrated into sentence descriptors give weighted accuracies (biased on the number of class samples) up to 85% with feature dimensionality of 3809 [30] and 1054 [16]. The latter are reported with a stratified cross-validation, i.e. each fold contains the same proportion of class samples introducing speaker dependencies. Also, no dev-set was used and system parameters were optimized on the test set. It is worth mentioning that in [16], UA dropped from 83.1% to 72.6% with leave-one-speaker-out instead of stratified cross-validation. Finally, a different approach combining a Universal Background Model (UBM) with Maximum a Posteriori (MAP) adaptation resulted in 81.35% accuracy [31]. Despite the better reported results, these indicate that our proposed features, much lower in original dimensionality (at most 340), evaluated with no speaker dependencies and using a held-out set, are promising especially when combined with complementary front-end descriptors and/or more powerful classifiers.

5. CONCLUSIONS AND FUTURE WORK

We proposed a new feature extraction framework for emotion classification based on modulation features computed with ESA. The inherent existence of amplitude and frequency modulation in speech is informative for our task, since the combination of modulation-spectral features almost always achieves improvement in EmoDB and AESI over the spectral-prosodic baseline. Results suggest that multi-band filtering can capture emotional information and denser sampling of frequency bands increases classification accuracy.

Future work will investigate the performance of modulation features in other emotional databases with spontaneous conversational speech. We will also combine the modulation features with established frameworks in emotion recognition, such as descriptors derived with OpenSmile [26]. AESI contains also video recordings enabling audiovisual fusion.

Acknowledgment: Most of this work was done at NTUA. P. Maragos' research work was supported by the project "COGNIMUSE" which is implemented under the "ARISTEIA" Action of the Operational Program Education and Lifelong Learning and is co-funded by the European Social Fund and Greek National Resources. The authors wish to thank Prof. C. Soldatos at the University of Athens for the help with the design and acquisition of the AESI database.

REFERENCES

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

[2] B. Schuller, S. Reiter, and G. Rigoll, "Evolutionary Feature Generation in Speech Emotion Recognition," in *Proc. ICME*, 2006, pp. 5–8.

[3] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "Combining Frame and Turn-Level Information for Robust Recognition of Emotions within Speech," in *Proc. Interspeech*, 2007, pp. 2249–52.

[4] C. C. Lee, Mower E., Busso C., S. Lee, and S. Narayanan, "Emotion Recognition Using a Hierarchical Binary Decision Tree Approach," *Speech Communication*, 53(6):1162–71, 2011.

[5] E. Mower, M.J. Matarić, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," *IEEE Trans. on Audio, Speech and Language Processing*, 19(5):1057–70, 2011.

[6] A. Metallinou, N. Katsamanis, and S. Narayanan, "Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information," *Image and Vision Computing*, 31(2):137–52, 2013.

[7] M. Wöllmer, B. Schuller, F. Eyben, and G. Rigoll, "Combining Long Short-Term Memory and Dynamic Bayesian Networks for Incremental Emotion-Sensitive Artificial Listening," *IEEE Journal Of Selected Topics in Signal Processing*, 4(5):867–81, 2010.

[8] F. Eyben, M. Wöllmer, and B. Schuller, "A multitask approach to continuous five-dimensional affect sensing in natural speech," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(1):1162–71, 2012.

[9] B. Schuller and G. Rigoll, "Timing Levels in Segment-Based Speech Emotion Recognition," in *Proc. Interspeech*, 2006, pp. 1818–21.

[10] G. Zhou, J.H.L Hansen, and J.F Kaiser, "Classification of speech under stress based on features derived form the nonlinear Teager energy operator," in *Proc. ICASSP*, 1998, pp. 549–52.

[11] S. Wu, T.H. Falk, and W.Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Communication*, 53(5):768–85, 2011.

[12] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE TSP*, 41(10):3024–51, 1993.

[13] D. Dimitriadis and P. Maragos, "Continuous Energy Demodulation Methods and Application to Speech Analysis," *Speech Comm.*, 48(7):819–37, 2006.

[14] D. Dimitriadis, P. Maragos, and A. Potamianos, "Robust AM-FM Features for Speech Recognition," *IEEE SPL*, 12(9):621–24, 2005.

[15] F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," in *Proc. Interspeech*, 2005, pp. 1517–20.

[16] A. Tawari and M. M. Trivedi, "Speech emotion analysis: exploring the role of context," *IEEE Transactions on Multimedia*, 12(6):502–9, 2010.

[17] D. Dimitriadis, P. Maragos, and A. Potamianos, "On the effects of filterbank design and energy computation on robust speech recognition," *IEEE TASLP*, 19(6):1504–16, 2011.

[18] L. Atlas and S. A. Shamma, "Joint acoustic and modulation frequency," *Journal on Applied Signal Processing*, pp. 668–75, 2003.

[19] P. Lieberman and B. M. Sheldon, "Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech," *J. Acoust. Soc. Am.*, 34(7):922–27, 1962.

[20] R. Plutchik, "The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *American Scientist*, 89(4):344–50, 2001.

[21] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, John Wiley & Sons, 2nd edition, 2000.

[22] S. J. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*, Cambridge, England, 2006.

[23] T. Joachims, B. Schölkopf, C. Burges, and A. Smola, *Making Large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning*, MIT-Press, 1999.

[24] A. Lazaridis, V. Bourna, and N. Fakotakis, "Comparative Evaluation of Phone Duration Models for Greek Emotional Speech," *Journal of Computer Science*, 6(3):341–49, 2010.

[25] D. Ververidis, I. Kotsia, C. Kotropoulos, and I. Pitas, "Multi-modal emotion-related data collection within a virtual earthquake emulator," in *LREC*, Morocco, 2008.

[26] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE - the Munich Versatile and Fast Open-Source Audio Feature Extractor," in *ACM Multimedia*, 2010.

[27] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, 1(6):80–83, 1945.

[28] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," in *Proc. ASRU*, 2009, pp. 552–57.

[29] B. Schuller and F. Burkhardt, "Learning with synthesized speech for automatic emotion recognition," in *Proc. ICASSP*, 2010, pp. 5150–53.

[30] S. Casale, A. Russo, G. Scebba, and S. Serrano, "Speech emotion classification using machine learning algorithms," in *Proc. ICSC*, 2008, pp. 158–65.

[31] I. Trabelsi, D. B. Ayed, and N. Ellouze, "Improved Frame Level Features and SVM Supervectors Approach for The Recogniton of Emotional States from Speech: Application to Categorical and Dimensional States," *IJIGSP*, 5(9):8–13, 2013.