Amplitude Modulation Features for Emotion Recognition from Speech

Md Jahangir Alam[1, 2], *Yazid Attabi*[2, 3], *Pierre Dumouchel*[3], *Patrick Kenny*[2], *D. O'Shaughnessy*[1]

[1] INRS-EMT, University of Quebec, Montreal (QC), Canada
[2] CRIM, Montreal (QC), Canada
[3]École de technologie supérieure, Montreal, Canada

{jahangir.alam,patrick.kenny,yazid.attabi}@crim.ca, pierre.dumouchel@etsmtl.ca, dougo@emt.inrs.ca

Abstract

The goal of speech emotion recognition (SER) is to identify the emotional or physical state of a human being from his or her voice. One of the most important things in a SER task is to extract and select relevant speech features with which most emotions could be recognized. In this paper, we present a smoothed nonlinear energy operator (SNEO)-based amplitude modulation cepstral coefficients (AMCC) feature for recognizing emotions from speech signals. SNEO estimates the energy required to produce the AM-FM signal, and then the estimated energy is separated into its amplitude and frequency components using an energy separation algorithm (ESA). AMCC features are obtained by first decomposing a speech signal using a *C*-channel gammatone filterbank, computing the AM power spectrum, and taking a discrete cosine transform (DCT) of the root compressed AM power spectrum. Conventional MFCC (Mel-frequency cepstral coefficients) and Mel-warped DFT (discrete Fourier transform) spectrum based cepstral coefficients (MWDCC) features are used for comparing the recognition performances of the proposed features. Emotion recognition experiments are conducted on the FAU AIBO spontaneous emotion corpus. It is observed from the experimental results that the AMCC features provide a relative improvement of approximately 3.5% over the baseline MFCC.

Index Terms: speech emotion recognition, amplitude modulation, SNEO, power function nonlinearity

1. Introduction

The goal of automatic recognition of human emotions from speech is to recognize the underlying emotional state of a speaker from his or her voice. Motivated by a broad range of commercially promising applications, speech emotion recognition has gained rapidly increasing research attention over the past few years [1]. In recent years a great deal of research has been done to automatically recognize emotions from human speech [1-10]. Some of this research has been further applied to call centers, multi-agent systems and other areas [11-15].

Extraction of features from a speech signal that efficiently characterize the emotional content of speech and at the same time do not depend on the speaker or lexical content is an important issue in speech emotion recognition [2, 16]. Speech signals may contain linguistic and paralinguistic features indicating emotional states. Thousands of paralinguistic features are extracted and used in experiments as a whole set or reduced to a subset using feature selection techniques. These features can be classified to one of three categories: Prosodic such as pitch (F0), intensity, and duration, Voice Quality such as jitter and shimmer, and Spectral such as MFCC (Mel-frequency cepstral coefficients) or LPCC (linear prediction cepstral coefficients) [6, 7, 16]. Among the features mentioned in the literature as being relevant for characterizing the manifestations of speech emotions, the most widely used are prosodic features. This is because the earliest studies of emotion detection were carried out using acted speech, where the linguistic content was controlled [16].

The spectral features, when used in combination with other categories of features (or even as a stand-alone feature vector), have been found to improve (or to achieve good) performance [6-7, 10, 17]. MFCC [18] and Perceptual Linear Prediction (PLP, with or without RASTA filtering) [19] are examples of spectral features that achieve good results not only on speech processing in general but also on emotion recognition [6-7, 9]. The problem with MFCC is that these features are easily affected by common frequency localized random perturbations, to which human perception is largely insensitive [8]. Since MFCC (and PLP) features are computed from windowed (e.g., Hamming) direct spectrum estimates (the squared magnitude of the Fourier transform of the observed signal), they suffer from high variance. Variance in the feature vectors has a direct bearing on the variance of Gaussians modeling the emotion classes. Reduction in the variance of the feature vector increases class separability and improved class separability can potentially increase recognition accuracy.

Amplitude modulation-frequency modulation (AM-FM) of a speech signal plays an important role in speech perception and recognition [8]. Recently, the AM-FM model has been applied in emotion recognition tasks [1, 8]. In [1], Hilbert transform based long-term (256 ms analysis frame length with a frame shift of 64 ms) modulation spectral features (MSFs) were proposed for emotion recognition. Teager energy operator (TEO)-based MFCC features (dubbed as TEMFCC and T-MFCC) were proposed in [8] for robust emotion recognition in noisy environments. In TEMFCC, TEO is applied on the complex DFT spectrum to compute the AM spectrum and other steps are the same as with the MFCC (see fig. 1(b)). In T-MFCC, TEO is applied on the windowed speech frame to separate AM components and then DFT is applied on the AM components; the rest of the steps are the same as the MFCC computation steps as shown in fig. 1(b).

A standard approach to the AM-FM demodulation problem is to use the Hilbert transform and the related Gabor's analytic signal [21]. An alternative approach is to use a nonlinear energy operator (NEO) to track the energy required to generate an AM-FM signal and separate it into amplitude and frequency components. The NEO approach to demodulation has many attractive features such as simplicity, efficiency, and

adaptability to instantaneous signal variations [20]. In this paper, we use smoothed nonlinear energy operator (SNEO) [21, 22] based amplitude modulation cepstral coefficients (AMCC) features for a speech emotion recognition task. The advantage of NEO is that it uses only a few samples of the input signal to estimate the energy required to generate an AM-FM signal and separate it into amplitude and frequency components without imposing any stationary assumption as done by linear prediction or Fourier transform [23]. A power function nonlinearity with a coefficient of 0.07 is applied, as it was found in [24] that it is more robust than the logarithmic nonlinearity used in conventional MFCC and MWDCC computation frameworks. The final features of the AMCC front-end are obtained by taking the Discrete Cosine Transform (DCT) and normalizing the features using the full utterance-based cepstral mean normalization method. Conventional MFCC and Mel-warped DFT spectrum [27]-based cepstral coefficients (MWDCC) are chosen for comparing the performance of the proposed features. Various steps involved in the extraction of the MFCC and MWDCC features are shown in fig. 1(b) and fig. 1(c), respectively. MFCC, MWDCC and AMCC features are used in experiments on the FAU AIBO corpus, a well-known spontaneous emotion speech corpus (described in section 3). The extracted features are used as short-term information (analysis frame length is 25 ms with a frame shift of 10 ms) and modeled using GMM models.

2. AMCC Feature Extraction

Extensive research by Teager resulted in a nonlinear approach for computing the energy of a signal denoted as the nonlinear energy operator (NEO) or Teager Kaiser energy operator (TKEO) [25]. The NEO uses only a few samples of the input signal to track the energy required to generate an AM-FM signal and separate it into amplitude and frequency components in a nonlinear manner, which provides an advantage over conventional Fourier transform (FT) or linear prediction (LP) methods in capturing the energy fluctuations. Fig. 1 (a) presents the block diagram of the proposed AMCC front-end. After removing the non-speech portions, pre-processing, framing and windowing, each frame of the windowed speech signal is decomposed using a *C*-channel (here, *C*=40 is used) gammatone filterbank covering the frequency range of 100-7200 Hz (for a sampling frequency of 16000 Hz). The AM power spectrum for each channel is then estimated using the smoothed nonlinear energy operator (SNEO).

Let $x(c,n)$ represent the speech frame of the cth channel, where $c = 1,2,...,C$ is the channel (or filterbank) index of the C-channel gammatone filterbank, $n = 1,2,...,N$ is the discrete time index, N is the frame length in samples and C is the number of channels of the gammatone filterbank. Standard NEO of $x(c,n)$ can be expressed as a special case of the following kth order ($k=0,1,2,...$) and lth lag ($l=1,2,3,...$) generalized discrete energy operator:

$$\Psi_{k,l}\left(x(c,n)\right) = x(c,n)x(c,n+k) - x(c,n-l)x(c,n+k+l). \quad (1)$$

For k=0 and l=1, eqn. (1) reduces to the standard NEO or TKEO. The NEO has the problem of cross terms and few negative values. To alleviate these problems we use the smoothed NEO (SNEO) [21, 22], which is expressed as:

$$\Psi^{s}_{0,1}\left(x(c,n)\right) = \Psi_{0,1}\left(x(c,n)\right) \otimes w(n), \quad (2)$$

where $w(n)$ is the smoothing window and $\otimes$ represents the convolution operator. For smoothing, a Bartlett window was used in [21], whereas in [20] a 7-point binomial smoothing filter with impulse response (1, 6.15, 20, 15, 6, 1) was applied. In this work we use the latter smoothing filter. Since NEO is an energy operator and energy is a positive quantity, in order to avoid any negative values in eqn. (2) (if $x(c,n)x(c,n+k) < x(c,n-l)x(c,n+k+l)$ for k=0, l=1) we have taken the absolute values of eqn. (1) [23, 25].

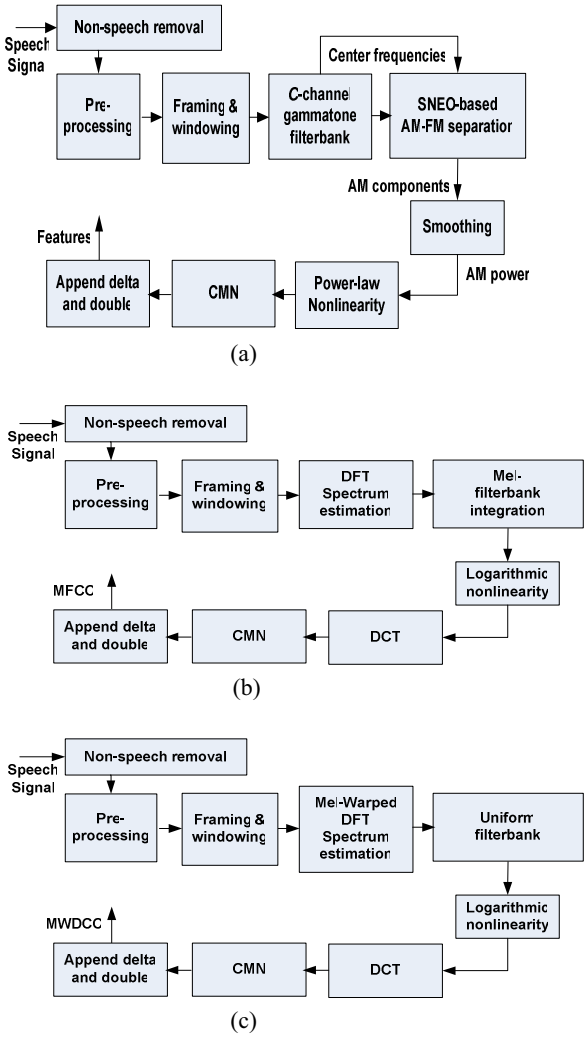

Figure 1: *Schematic diagrams showing (a) the various steps of the amplitude modulation cepstral coefficients (AMCC) feature for automatic speech emotion recognition, (b) various steps of the baseline MFCC feature extraction process, and (c) Mel-warped DFT spectrum based MFCC features (MWDCC) computation framework.*

For the *c*th channel, the AM and FM components can be estimated using the discrete energy separation algorithm (DESA) when k=0, l=1 as follows:

$$\left|\hat{a}(c,n)\right| = \sqrt{\frac{\Psi^s_{0,1}\left(x(c,n)\right)}{1-\left(1-\frac{\Psi^s_{0,1}\left(y(c,n)\right)+\Psi^s_{0,1}\left(y(c,n+1)\right)}{4\Psi^s_{0,1}\left(x(c,n)\right)}\right)^2}}, \quad (3)$$

$$\left|\hat{\varphi}(c,n)\right| = \cos^{-1}\left(1-\left(\frac{\Psi^s_{0,1}\left(y(c,n)\right)+\Psi^s_{0,1}\left(y(c,n+1)\right)}{4\Psi^s_{0,1}\left(x(c,n)\right)}\right)\right), \quad (4)$$

where $y(c,n)) = x(c,n)) - x(c,n-1))$. In order to reduce the dynamic range, estimated AM and FM components are smoothed using a moving average filter with a window size of 5. In fig. 2, a windowed speech frame and its corresponding AM components with and without smoothing are presented. For the *m*th speech frame the AM power for the *c*th channel is computed as:

$$P(m,c) = \sum_{n=1}^{N}\left(\left|\hat{a}(c,n)\right|^2\right). \quad (5)$$

The 13-dimensional static features, obtained after applying a power function nonlinearity, using a coefficient of 0.07 and the discrete cosine transform (DCT), are normalized using the conventional cepstral mean normalization method. Finally, 39-dimensional features are obtained by computing the delta and double delta features with a 5-frame window and appending with the static features.

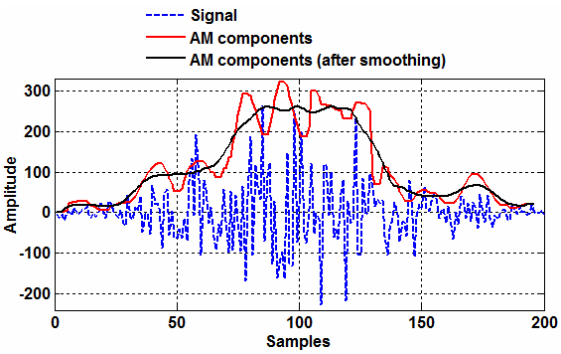

Figure 2: *A frame of a speech signal and its corresponding AM components with and without smoothing.*

3. Performance Evaluation

3.1. Emotion Recognition Corpus

The effectiveness of the AMCC features on an emotion recognition task is tested using the FAU AIBO [17] emotional speech corpus. The dataset consists of spontaneous recordings of German children interacting with a pet robot. The corpus is composed of 9959 chunks for training and 8257 chunks for testing. A chunk is an intermediate unit of analysis between the word and the turn, which is manually defined based on syntactic-prosodic criteria. The chunks are labeled into five emotion categories: *Anger* (**A**), *Emphatic* (**E**), *Neutral* (**N**), *Positive* (**P**, composed of *motherese* and *joyful*) and *Rest* (**R**, consisting of emotions not belonging to the other categories such as *bored*, *helpless*, and so on). The distribution of the five classes is highly unbalanced. For example, the percentage of training data of each class is as follows: **A**(8.8%), **E**(21%), **N**(56.1%), **P**(6.8%), **R**(7.2%).

3.2. Gaussian Mixture Models (GMMs)

Cepstral feature vectors are modeled using a GMM model. GMM is a generative model widely used in the field of speech processing. It is a semi-parametric probabilistic method that offers the advantage of adequately representing speech signal variability. Given a GMM modeling a *D*-dimensional vector, the probability of observing a feature vector given the model λ is computed as follows:

$$P(\mathbf{x}|\lambda) = \sum_{k=1}^{m} w_k N(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad (6)$$

where m, w_i, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ correspond to the number of Gaussians, weight, mean vector and diagonal covariance matrix of the *k*th Gaussian, respectively.
GMM parameters are estimated using a Maximum Likelihood (ML) approach based on the Expectation Maximization (EM) algorithm [28]. The classification of a test sequence of frames $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ is based on the Bayes decision. Using an equal prior probability for all classes, the classification is achieved by computing the log-likelihood of the test utterance against the GMM of each emotion class. The test recording is classified as the emotion class label that maximizes the log-likelihood value over all class models [7].

3.3. Experimental setup

The training of GMM models has been made with different numbers of mixtures taken from the set {2,4,8,16,32,64,128,256,512,1024}. The best parameter is tuned based on the training data using a 9-fold cross validation protocol, i.e., the number of components are tuned using 9-fold cross validation on the training set to maximize the performance separately for each system. Each fold contains a separate group of speakers to ensure speaker independent evaluation. After optimization, the selected numbers of Gaussians used for test data are as follows: 64 for the baseline MFCC, 128 for MWDCC and 256 for the AMCC systems.

The metrics used for the evaluation of automatic speech emotion recognition (ASER) performances are: unweighted average recall (UAR) and weighted average recall (WAR). WAR measure is preferred when the classes of the studied corpora is quite balanced. If the distribution of the classes in the studied corpus (e.g., FAU AIBO corpus) are very unbalanced then UAR measure is usually used for performance evaluation.

The results are optimized to maximize the unweighted average recall (UAR) measure and secondly the weighted average recall (WAR) (namely accuracy) given that FAU AIBO emotion classes are highly unbalanced (i.e., one class is disproportionately more represented than the others). Note that a baseline classifier that predicts all the test data of the same

class as of the majority one, namely *Neutral*, will achieve 65% of accuracy but only 20% of UAR.

3.4. Results and discussion

Table 1 presents the results obtained using the baseline systems and the proposed AMCC system. It is observed from this table that the AMCC system outperformed the baseline MFCC and MWDCC systems in terms of both UAR and WAR measures. Experimental results show that the MFCC obtained via the direct warping of DFT (discrete Fourier transform) spectrum [27], denoted in this work as MWDCC, also achieved better recognition accuracy (relative improvement is 9.1%), in terms of WAR scoring metric, than the conventional MFCC. The performance of MWDCC is almost same as the MFCC in UAR scoring metric. Relative improvements obtained by the AMCC, in UAR metric, over the baseline MFCC and MWDCC is approximately 3.2% and 3.27%, respectively. In WAR metric, the relative improvements are approximately 12.0% and 2.65% over the MFCC and MWDCC, respectively. Our future work is compare the performance of the proposed feature with that of T-MFCC & TEMFCC and to do fusion of different types of features (spectral and prosodic features) for emotion recognition.

Table 1. *Emotion recognition results achieved on FAU AIBO test data for the baseline MFCC and MWDCC (Mel-warped DFT spectrum-based MFCC) and proposed AMCC systems in terms of UAR and WAR scoring metrics.*

	UAR (%)	**WAR (%)**
MFCC	43.09	37.99
MWDCC	43.11	41.48
AMCC	**44.50**	**42.58**

4. Conclusions

In this paper, we present amplitude modulation features-based cepstral coefficients (AMCC) for emotion recognition from speech. AMCC features are derived from a short-term spectro-temporal representation of a speech signal via a smoothed nonlinear Teager energy operator (SNEO). Performance of the proposed features is compared, in the context of speech emotion recognition task on the FAU AIBO emotion corpus, with the baseline MFCC and MWDCC systems. Proposed features are shown to outperform the baseline features in terms of emotion recognition accuracy measured using UAR and WAR scoring metrics.

5. References

[1] Siqing Wu, Tiago H. Falk, Wia-Yip Chan, "Automatic speech emotion recognition using modulation spectral features," Speech Comm., vol. 53, no. 5, pp. 768-785, 2011.

[2] Lijiang Chen, Xia Mao, Yuli Xue, Lee Lung Cheng, "Speech emotion recognition: Features and classification models," Digital Signal Processing, vol. 22, pp. 1154-1160, 2012.

[3] D. Ververidis, C. Kotropoulos, "Emotional speech recognition – resources features and methods," Speech Commun. vol. 48, pp. 1162–1181, 2006.

[4] K. Scherer, "Vocal communication of emotion: A review of research paradigms," Speech Commun. vol. 40, pp. 227–256, 2003.

[5] T. Sobol-Shikler, P. Robinson, "Classification of complex information: Inference of co-occurring affective states from their expressions in speech," IEEE Trans. Pattern Anal. Mach. Intell. vol. 32, no. 7, pp. 1284–1297, 2010.

[6] P. Dumouchel, N. Dehak, Y. Attabi, R. Dehak, N. Boufaden: "Cepstral and long-term features for emotion recognition," *Proc. INTERSPEECH* , pp. 344-347, 2009.

[7] Y. Attabi, M. J. Alam, P. Dumouchel, P. Kenny, D. O'Shaughnessy, "Multiple windowed spectral features for emotion recognition," Proc. ICASSP (to appear), Vancouver, 2013.

[8] Alexandros Georgogiannis, Vassilis Digalakis, "Speech emotion recognition using nonlinear Teager energy based features in noisy environments," Proc. EUSIPCO, Bucharest, Romania, August, 2012.

[9] Nobuo Sato, Yasunari Obuchi, "Emotion recognition using Mel-frequency cepstral coefficients," Journal of Natural Language Processing, vol. 14, no. 4, pp. 83-96, 2007.

[10] Neiberg D, Elenius K, Laskowski K, "Emotion recognition in spontaneous speech using GMMs," *Proc. of INTERSPEECH Conference*, pp 809–812, 2006.

[11] C. Peter, Affect and Emotion in Human–Computer Interaction: From Theory to Applications, vol. 4868, Springer-Verlag, New York, 2008.

[12] W. Yoon, K. Park, "A study of emotion recognition and its applications," Modeling Decisions for Artificial Intelligence, vol. 6417, pp. 455–462, 2007.

[13] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, H. Konosu, "Being bored? Recognizing natural interest by extensive audiovisual integration for real-life application," Image Vis. Comput. vol. 27, no. 12, pp. 1760-1774, 2009.

[14] K. Van Deemter, B. Krenn, P. Piwek, M. Klesen, M. Schröder, S. Baumann, "Fully generated scripted dialogue for embodied agents," Artificial Intelligence, vol. 172, no. 10, pp. 1219–1244, 2008.

[15] E. Lorini, F. Schwarzentruber, "A logic for reasoning about counterfactual emotions," Artificial Intelligence, vol. 175, no. 3, pp. 814–847, 2011.

[16] Klaus R. Scherer, Tanja Bänziger, Etienne B. Roesch, Edited book, Blueprint for Affective Computing - A Sourcebook, Oxford University press, Oxford, NY, 2010.

[17] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," *Interspeech*, ISCA, Brighton, UK, 2009.

[18] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 28(4), pp. 357–366, 1980.

[19] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87 (4), pp. 1738-1752, 1990.

[20] Maragos, J. F. Kaiser, and T. F. Quatieri, On amplitude and frequency demodulation using energy operators, IEEE Trans. Signal Processing , vol. 41, no. 4, pp. 1532–1550, April 1993.

[21] Alexandros Potamianos, Speech processing applications using an AM-FM modulation model, PhD thesis, Harvard University, Cambridge, MA, 1995.

[22] S. Mukhopadhyay, G. C. Ray, A new interpretation of nonlinear energy operator and its efficacy in spike detection, IEEE Tans. on Biomedical Engg., vol. 45, no. 2, pp. 180-187, 1998.

[23] Vikramjit Mitra, H. Franco, M. Graciarena, A. Mandal, Normalized Amplitude modulation features for large vocabulary noise-robust speech recognition, Proc. of ICASSP, pp. 4117-4120, 2012.

[24] C. Kim and R. M. Stern., Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring, In IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, pp. 4574-4577, March 2010.

[25] J.F. Kaiser, On a Simple Algorithm to Calculate the 'Energy' of a Signal," *Proceedings of IEEE International Conference on*

Acoustics, Speech, and Signal Processing, Albuquerque, NM, pp. 381-384, Apr. 1990.

[26] Eivind Kvedalen, Signal processing using the Teager energy operator and other nonlinear operators, Cand. Scient thesis, University of Oslo, May 2003.

[27] Umit H. Yapanel, John H.L. Hansen, "A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition," Speech Comm., Vol. 50, pp. 142-152, 2008.

[28] A. P. Dempster, N. M. laird, and. D. B. Robin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royel Stastical Society*, vol. B, pp. 1-38, 1997.