Automatic Emotion Recognition from Cochlear Implant-Like Spectrally Reduced Speech

Md Jahangir Alam [1], Yazid Attabi[1,2], Patrick Kenny [1], Pierre Dumouchel[2],
and Douglas O'Shaughnessy [3]

[1] CRIM, Montreal (QC) Canada
[2] ETS, Montreal (QC) Canada
[3] INRS-EMT, University of Quebec, Montreal (QC) Canada
{jahangir.alam,yazid.attabi,patrick.kenny}@crim.ca

Abstract. In this paper we present a robust feature extractor that includes the In this paper we study the performance of emotion recognition from cochlear implant-like spectrally reduced speech (SRS) using the conventional Mel-frequency cepstral coefficients and a Gaussian mixture model (GMM)-based classifier. Cochlear-implant-like SRS of each utterance from the emotional speech corpus is obtained only from low-bandwidth subband temporal envelopes of the corresponding original utterance. The resulting utterances have less spectral information than the original utterances but contain the most relevant information for emotion recognition. The emotion classes are trained on the Mel-frequency cepstral coefficient (MFCC) features extracted from the SRS signals and classification is performed using MFCC features computed from the test SRS signals. In order to evaluate to the performance of the SRS-MFCC features, emotion recognition experiments are conducted on the FAU AIBO spontaneous emotion corpus. Conventional MFCC, Mel-warped DFT (discrete Fourier transform) spectrum-based cepstral coefficients (MWDCC), PLP (perceptual linear prediction), and amplitude modulation cepstral coefficient (AMCC) features extracted from the original signals are used for comparison purpose. Experimental results depict that the SRS-MFCC features outperformed all other features in terms of emotion recognition accuracy. Average relative improvements obtained over all baseline systems are 1.5% and 11.6% in terms of unweighted average recall and weighted average recall, respectively.

Keywords: Automatic emotion recognition, cochlear implant, spectrally reduced speech, MFCC, AMCC, GMM.

1 Introduction

The aim of automatic emotion recognition (AER) 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].

L. Pecchia et al. (Eds.): IWAAL 2014, LNCS 8868, pp. 332–340, 2014.

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. The paralinguistic 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]. Reduction of speech variability due speech production or environment is important to achieve robust emotion recognition performances. Therefore, in an AER system, the aim of speech analysis module is to reduce signal variability and extract relevant acoustic features for emotion recognition. In spite of speech variability reduction achieved by the standard MFCC and PLP features AER performance is still affected by the sources of speech variability. As most of the emotion recognition features are extracted by analyzing speech in the spectral domain it is natural to seek the relevant spectral information from the speech signal that is sufficient for AER [20]. One technique to estimate relevant speech spectral information for a GMM (gaussian mixture model)-based AER system is to train GMM models for emotion classes and evaluate emotion recognition performance on the cochlear implant-like spectrally reduced speech (SRS) signals. The acoustic simulation of a cochlear implant is a spectrally reduced transform of original speech and it has been shown in [24] that normal hearing listeners could achieve a nearly perfect recognition score when listening to these SRS signals.

MFCC and PLP front-end, which mimic the speech processing performed by the human auditory system, are basically aimed at reducing the acoustic variability while putting emphasis on the most relevant spectral information for recognition. Therefore, cochlear implant-like SRS should contain sufficient information for AER based on conventional MFCC or PLP features. Inspired from the algorithm, introduced in [24], to synthesize acoustic simulation of a cochlear implant, spectrally reduced speech has already been applied in HMM-based automatic speech recognition [20-22], and GMM-UBM -based speaker verification [23] tasks. In this work, our objective is to find out whether cochlear implant-like SRS contains sufficient spectral information for AER based on the conventional MFCC features.

In order to evaluate the performance of SRS-MFCC features and make a comparison with the original speech-based cepstral features MFCC, PLP, MWDCC (Mel-warped DFT spectrum-based cepstral coefficients), AMCC (amplitude modulation cepstral coefficient), and SRS-MFCC features are used in experiments on the FAU AIBO corpus, a well-known spontaneous emotion speech corpus. 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. Experimental results show the effectiveness of the SRS-MFCC features in terms of emotion recognition accuracy.

2 Cepstral Features from Spectrally Reduced Speech

This section describes the procedure to obtain spectrally reduced speech (SRS) from the original speech and compute mel-frequency cepstral coefficients (MFCC) features from it. Here, we denote this as SRS-MFCC features. Fig. 1 presents a complete block diagram for the SRS-MFCC feature extraction process and Fig. 2 shows the various steps to compute MFCC features from the SRS signal.

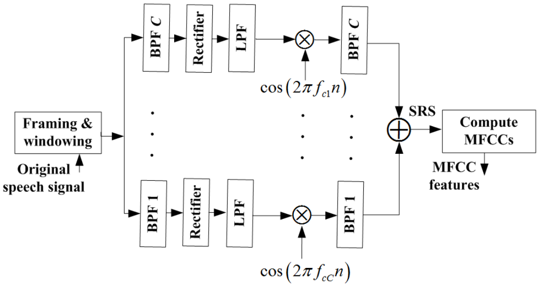

Fig. 1. Block diagram showing various steps to obtain spectrally reduced speech (SRS) from a original speech signal and then computation of mel-frequency cepstral coefficients (MFCC) features from that SRS signal. BPF and LPF stands for bandpass filter and low-pass filter, respectively.

Original speech signal is first framed (frame length is 20 ms with a frame shift of 10 ms) and windowed using a Hamming window. The windowed speech signal is then decomposed into C channel (or subband) signals $x_{c'}(t), c'=1,2,...,C$ by applying a perceptually motivated analysis filterbank and overlap-add technique. The analysis filterbank consists of C non-uniform bandwidth bandpass filters (BPFs) which are linearly spaced on the Bark scale in order to approximate the nonlinear characteristics of the human auditory system. Each bandpass filter (BPF) in the filterbank is a 2nd order elliptic BPF having a minimum stopband attenuation of 50 dB and a 2 dB peak-to-peak ripple in the passband [20, 23]. The lower, upper, and central frequencies of the BPFs are computed in the same way as described in [27]. Fig. 3 presents the frequency response of an analysis filterbank comprised of $C=16$ second order BPFs that are linearly spaced on the Bark scale.

The c'-th channel amplitude modulation $a_{c'}(t)$ (or temporal envelope) of the c'-th signal $x_{c'}(t), c'=1,2,...,C$ is then obtained by applying a low pass filter followed by full-wave rectification of the output signal of the c'-th channel bandpass filter. The purpose of using a low-pass filter, a fourth order elliptic LPF with 2 dB of peak-to-peak ripple and a minimum stopband attenuation of 50 dB, is to limit the bandwidth of the subband temporal envelopes.

The c'-th channel amplitude modulation $a_{c'}(t)$ is then used to modulate a sinusoid whose frequency f equals the centre frequency $f_{cc'}$ of the BPF of that channel.

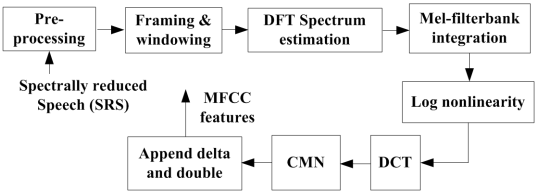

Fig. 2. Various steps for the MFCC feature extraction process from the spectrally reduced speech signals

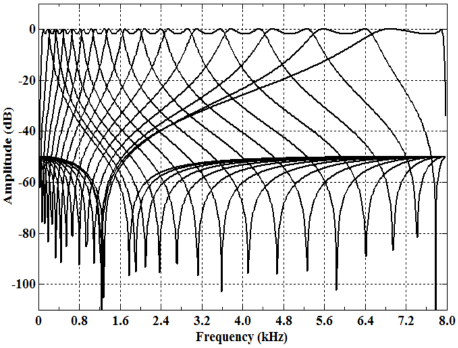

Fig. 3. Frequency response of an analysis filterbank consisting of sixteen 2nd order elliptic bandpass filters (BPFs) that are linearly spaced on the Bark scale. Sampling frequency is 16 kHz.

The modulated signal of c'-th channel $\hat{x}_{c'}(t) = a_{c'}(t)\cos(2\pi f_{cc'}t)$ is again bandpass filtered using the same BPF used for the original analysis subband. If $\mathbb{F}_{BPF}^{c'}(\cdot)$ denotes the c'-th channel bandpass filtering operation then the spectrally reduced signal (SRS) $\hat{x}(t)$ of the original signal can be expressed as:

$$\begin{aligned}\hat{x}(t) &= \sum_{c'=1}^{C} \mathbb{F}_{BPF}^{c'}\left(\hat{x}_{c'}(t)\right) \\ &= \sum_{c'=1}^{C} \mathbb{F}_{BPF}^{c'}\left(a_{c'}(t)\cos\left(2\pi f_{cc'}t\right)\right).\end{aligned} \tag{1}$$

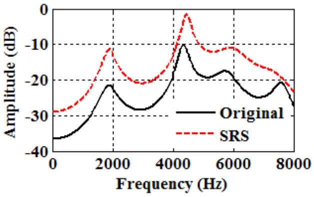

Fig. 4. All-pole spectral envelopes (using linear prediction with a model order of 20) of a frame of original speech and the corresponding spectrally reduced speech (SRS). Sampling frequency of the speech signal is 16 kHz. Number of subbands in the analysis filterbank is 16 and the cutoff frequency of the LPF (low-pass filter) is 50 Hz.

Fig. 4 shows short-term spectral envelopes of a frame of original speech (taken from the emotion corpus) and the corresponding SRS signals. Linear prediction with a model order of $p = 20$ is used to estimate the short-term all-pole spectral envelopes. Fig. 4 demonstrates that the global shapes of the all-pole spectral envelope of the SRS signal, obtained with C = 16, and $f_c = 50$ Hz, frame is rather similar (specifically, up to 6 kHz) with that of the all-pole spectral envelope of that frame of original speech. By increasing the number of subbands C and the cut off frequency f_c it is possible to obtain SRS spectral envelopes that are more similar to the original speech spectral envelopes [23].

The SRS signal it is then passed through the feature extraction process to compute cepstral features. MFCC processing begins with pre-emphasis, typically using a first-order high-pass filter. Short-time Fourier Transform (STFT) analysis is performed using a hamming window, and triangular-shaped Mel-frequency integration is performed for auditory spectral analysis. The logarithmic nonlinearity stage follows, and the 13-dimensional static features are obtained through the use of a Discrete Cosine Transform (DCT). After normalizing the static features using a cepstral mean normalization (CMN) technique, first and second derivatives are appended with the static features, making a final set of 39-dimensional MFCC features.

3 Emotion Recognition Experiments

The effectiveness of the spectrally reduced speech-mel-frequency cepstral coefficients (SRS-MFCC) features on an emotion recognition task is tested using the FAU AIBO [28, 17] emotional speech corpus. For comparison the following features computed from the original signal are chosen: conventional MFCC [18], Mel-warped DFT (discrete Fourier transform) spectrum-based cepstral coefficients (MWDCC) [7], and amplitude modulation cepstral coefficients (AMCC) [7] features. The dimension of features for each system is d = 39 and all systems use the cepstral mean normalization method as a post-processing scheme to normalize the static features.

3.1 Emotion Recognition Corpus

The FAU AIBO 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 $M = \{w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$ is computed as follows:

$$P(\mathbf{x}|M) = \sum_{i=1}^{m} w_i N(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \quad (2)$$

where m, w_i $\boldsymbol{\mu}_i$, and $\boldsymbol{\Sigma}_i$ correspond to the number of Gaussians, weight, mean vector and diagonal covariance matrix of the *i*-th Gaussian, respectively.

GMM parameters are estimated using a Maximum Likelihood (ML) approach based on the Expectation Maximization (EM) algorithm [26]. The classification of a test sequence of *T* 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 separately for each system based on the training data using a 9-fold cross validation protocol. 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: 128 for the baseline MFCC, 128 for MWDCC, 128 for PLP, 256 for the AMCC, and 256 for SRS-MFCC systems. The metrics used for the evaluation of automatic speech emotion recognition performances are: unweighted average recall (UAR) and weighted average recall (WAR). The results are optimized to maximize the UAR measure and secondly the WAR (namely accuracy) given that FAU AIBO emotion classes are highly unbalanced (i.e., one class is disproportionately more represented than the others).

3.4 Results and Discussion

Similar to [20-23, 24], in order to evaluate the effect of reducing the bandwidth of the temporal envelope information we did emotion recognition experiments by varying the cut-off frequency f_c of the low-pass filter (LPF) from 16 Hz to 500 Hz. The value of f_c was chosen optimal that provided highest emotion recognition accuracy. To find the optimal number of subbands, we synthesized spectrally reduced speech (SRS) from the original speech by varying the number of subbands (or channels) *C* from 16 to 50 and found that C = 16 with $f_c = 50$ Hz provided highest accuracy. Here, we report emotion recognition results on the eval (or test) data for C = 16 & $f_c = 50$ Hz.

Table 1 presents the results obtained using the baseline systems and the SRS-MFCC system. It is observed from this table that the SRS-MFCC system outperformed the baseline MFCC, PLP and MWDCC systems in terms of both UAR and WAR measures. Although the performance of SRS-MFCC is close to that of AMCC in terms of the UAR metric SRS-MFCC outperformed AMCC in WAR metric. It has been shown in [7] that the MFCC obtained via the direct warping of the DFT (discrete Fourier transform) spectrum, denoted as MWDCC, achieved better recognition accuracy, in terms of the WAR scoring metric, than the conventional MFCC. The performance of MWDCC was almost the same as the MFCC in UAR scoring metric. Relative improvements obtained by the SRS-MFCC, in UAR metric, over the baseline MFCC, PLP, MWDCC, and AMCC are approximately 1.3%, 3.9%, 1.9% and -1.3%, respectively. With the WAR metric, the relative improvements are approximately 14.9% and 11.4%, 11.5%, and 8.6%, over the MFCC, PLP, MWDCC, and AMCC, respectively. Presented results demonstrate that the cochlear implant-like SRS is a relevant speech model for using in AER. Our future work is to compute PLP and AMCC features from SRS signals and compare their performances with the PLP and AMCC features computed from the original speech signals.

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

	UAR (%)	**WAR (%)**
MFCC	43.37	40.26
PLP	42.30	41.50
MWDCC	43.11	41.48
AMCC	**44.50**	42.58
SRS-MFCC	43.94	**46.24**

4 Conclusion

In this paper, we present spectrally reduced speech (SRS) -based Mel-frequency cepstral coefficients (SRS-MFCC) features for emotion recognition. Inspired from speech signal processing algorithms in standard cochlear implants, the SRS signals are obtained by applying cochlear implant-like synthesis algorithm to the original emotion corpus. Although SRS has reduced spectral information than the original one it is observed, experimentally, that SRS-MFCC features carry relevant information for emotion recognition. Performance of the SRS-MFCC features is compared, in the context of speech emotion recognition task on the FAU AIBO emotion corpus, with the conventional MFCC, PLP, MWDCC, and AMCC systems. SRS-MFCC features are shown to outperform the baseline features in terms of emotion recognition accuracy measured using UAR and WAR scoring metrics. Average relative improvements obtained over all baseline systems are 1.5% and 11.6% in terms of UAR and WAR, respectively.

References

[1] Wu, S., Falk, T.H., Chan, W.-Y.: Automatic speech emotion recognition using modulation spectral features. Speech Comm. 53(5), 768–785 (2011)

[2] Chen, L., Mao, X., Xue, Y., Cheng, L.L.: Speech emotion recognition: Features and classification models. Digital Signal Processing 22, 1154–1160 (2012)

[3] Ververidis, D., Kotropoulos, C.: Emotional speech recognition – resources features and methods. Speech Commun. 48, 1162–1181 (2006)

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

[5] Sobol-Shikler, T., Robinson, P.: Classification of complex information: Inference of co-occurring affective states from their expressions in speech. IEEE Trans. Pattern Anal. Mach. Intell. 32(7), 1284–1297 (2010)

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

[7] Alam, M.J., Attabi, Y., Dumouchel, P., Kenny, P., O'Shaughnessy, D.: Amplitude Modulation Features for Emotion Recognition from Speech. In: Proc. INTERSPEECH, Lyon, France (2013)

[8] Georgogiannis, A., Digalakis, V.: Speech emotion recognition using nonlinear Teager energy based features in noisy environments. In: Proc. EUSIPCO, Bucharest, Romania (August 2012)

[9] Sato, N., Obuchi, Y.: Emotion recognition using Mel-frequency cepstral coefficients. Journal of Natural Language Processing 14(4), 83–96 (2007)

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

[11] Peter, C., Beale, R. (eds.): Affect and Emotion in Human-Computer Interaction. LNCS, vol. 4868. Springer, Heidelberg (2008)

[12] Yoon, W.-J., Park, K.-S.: A study of emotion recognition and its applications. In: Torra, V., Narukawa, Y., Yoshida, Y. (eds.) MDAI 2007. LNCS (LNAI), vol. 4617, pp. 455–462. Springer, Heidelberg (2007)

[13] Schuller, B., Müller, R., Eyben, F., Gast, J., Hörnler, B., Wöllmer, M., Rigoll, G., Höthker, A., Konosu, H.: Being bored? Recognizing natural interest by extensive audiovisual integration for real-life application. Image Vis. Comput. 27(12), 1760–1774 (2009)

[14] Van Deemter, K., Krenn, B., Piwek, P., Klesen, M., Schröder, M., Baumann, S.: Fully generated scripted dialogue for embodied agents. Artificial Intelligence 172(10), 1219–1244 (2008)

[15] Lorini, E., Schwarzentruber, F.: A logic for reasoning about counterfactual emotions. Artificial Intelligence 175(3), 814–847 (2011)

[16] Scherer, K.R., Bänziger, T., Roesch, E.B. (eds.): Blueprint for Affective Computing - A Sourcebook. Oxford University Press, Oxford (2010)

[17] Schuller, B., Steidl, S., Batliner, A.: The INTERSPEECH 2009 Emotion Challenge. In: Interspeech, ISCA, Brighton (2009)

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

[19] Hermansky, H.: Perceptual linear predictive (PLP) analysis of speech. Journal of the Acoustical Society of America 87(4), 1738–1752 (1990)

[20] Do, C.-T., Pastor, D., Goalic, A.: A novel framework for noise robust ASR using cochlear implant-like spectrally reduced speech. Speech Communication 54(1), 119–133 (2012)

[21] Do, C.-T., Pastor, D., Le Lan, G., Goalic, A.: Recognizing cochlear implant-like spectrally reduced speech with HMM-based ASR: experiments with MFCCs and PLP coefficients. In: Proc. of INTERSPEECH 2010, pp. 2634–2637 (September 2010)

[22] Do, C.-T., Taghizadeh, M.J., Garner, P.N.: Combining cepstral normalization and cochlear implant-like speech processing for microphone array-based speech recognition. In: Proc. SLT 2012 - IEEE Workshop on Spoken Language Technology, pp. 137–142 (December 2012)

[23] Do, C.-T., Barras, C.: Cochlear implant-like processing of speech signal for speaker verification. In: Proc. SAPA 2012 Conference - Statistical and Perceptual Audition (Satellite Workshop of Interspeech 2012), pp. 17–21 (September 2012)

[24] Shannon, R.V., Zeng, F.-G., Kamath, V., Wygonski, J., Ekelid, M.: Speech recognition with primarily temporal cues. Science 270(5234), 303–304 (1995)

[25] Zeng, F.-G., Nie, K., Stickney, G., Kong, Y.-Y., Vongphoe, M., Bhargave, A., Wei, C., Cao, K.: Speech recognition with amplitude and frequency modulations. Proceedings of National Academy of Sciences 102(7), 2293–2298 (2005)

[26] Dempster, A.P., Laird, N.M., Robin, D.B.: Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royel Stastical Society B, 1–38 (1997)

[27] Gunawan, T.S., Ambikairajah, E.: Speech enhancement using temporal masking and fractional Bark gammatone filters. In: Proc. 10th Australian Int. Conf. Speech Sci. Technol., Sydney, Australia, December 08-10, pp. 420–425 (2004)

[28] Steidl, S.: Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech. Logos Verlag, Berlin (2009)