A Novel Feature Extraction Strategy for Multi-stream Robust Emotion Identification

Gang Liu, Yun Lei, John H. L. Hansen

CRSS: Center for Robust Speech Systems
Erik Jonsson School of Engineering and Computer Science
University of Texas at Dallas, Richardson, Texas 75083, USA
{gxl083000, yxl059200, John.Hansen}@utdallas.edu

Abstract

We investigate an effective feature extraction front-end for speech emotion recognition, which performs well in clean and noisy conditions. First, we explore the use of perceptual minimum variance distortionless response (PMVDR). These features, originally proposed for accent/dialect and language identification (LID), can better approximate the perceptual scales and are less sensitive to noise and speaker variation. Also developed for LID, shifted delta cepstral (SDC) approach can be used to incorporate additional temporal information. It is known that supra-segmental speech characteristics, such as pitch and intensity, provide better discriminative information for emotion recognition by fusing with other emotion dependent features. Combined PMVDR and SDC together, the system outperforms the baseline system (MFCC based) by 10.3% (absolute). Furthermore, we find both PMVDR and SDC offer much better robustness in noisy condition, which is critical for real applications. All the evaluation the proposed features using the Berlin database of emotion speech.

Index Terms: PMVDR, shifted delta cepstral, emotion identification, robustness

1. Introduction

Reliable emotion identification/reocgintion (this two terms are used interchangeably) can be used to increase the performance of speech/speaker recognition systems for a range of applications such as spoken dialog systems, human computer interfaces, and gaming industry. Recently, extensive research has been conducted in the speech area to improve the performance of the stress/emotion recognition [1]-[6].

In the emotion identification (EID) field, various types of features have been proposed, such as Mel-frequency cepstrum coefficients (MFCCs), pitch, intensity, and their variation [7, 8]. One problem with EID systems is that it usually involves huge dimension feature sets. For example, more than 200 features were used in [7] and about 1000 features were used in [8]. The feature sets are usually reduced with feature selection algorithms or principle component analysis. As a result, the feature extraction process is computation demanding and the selected features are highly dependent on the emotional database and are not robust against noisy signals. In this study, we focus on providing an effective and robust feature extraction front-end for emotion recognition system, which only utilizes few robust features that have the potential to perform well in online and real life task.

In addition to features that are commonly used for emotion recognition, we propose the use of perceptual minimum variance distortionless response (PMVDR) features and shifted delta cepstrum (SDC). MFCCs [9] have proven to be one of the most effective feature sets for speech processes, especially automatic speech recognition (ASR) and speaker identification. They are computed by applying a Mel-scaled filter-bank either to the short-term FFT magnitude spectrum or to the short-term LPC-based spectrum to obtain a perceptually meaningful smoothed gross spectrum. Both the FFT and LPC-based spectrum, however, have limited ability to remove undesired harmonic structures, especially for high pitch speech [10], which may affect emotion representation. Furthermore, studies have shown that FFT-based MFCCs are less effective for stressed speech recognition than LP-based MFCCs [11]. Moreover, MFCCs are expected to carry speaker dependent information. In fact, MFCCs are commonly used by almost all the state-of-the-art NIST SRE08 systems, including the best one [12]. In contrast, PMVDR directly warps the FFT power spectrum of speech during the feature estimation process, removing the traditional Mel-scaled filterbank as a perceptually motivated frequency partitioning. It can provide a better approximation of the perceptual scales. Another advantage is that PMVDR can effectively model medium and high-pitch speech and track the upper envelope. Therefore, it smoothes undesired speaker excitation information to better suppress speaker dependent information and yields more accurate recognition and faster emotion decoding in both clean and noisy conditions.

In language identification, shifted delta cepstrum (SDC) approach [13] is widely used. The motivation to include this in the context of emotion recognition is to incorporate additional temporal information into the feature vector.

The remainder of this paper is organized as follows: Sec. 2 describes the database that is used to develop and evaluate the system. Sec. 3 describes the experiment baseline system. Sec. 4 presents the different feature extraction schemes. Sec. 5 discusses the EID results obtained from different feature extraction front-end schemes, and Sec. 6 presents conclusion and future work.

2. Corpus

This study employs data from acted emotional speech, the Berlin Emotional database (EMO-DB) [14]. Although acted material has a number of well-known drawbacks, many studies have used this corpus (is a benchmark database for EID). Here, we decide to use it to establish a proof of concept for the proposed methodology. EMO-DB contains seven categories of emotions (Table 1). Ten German sentences of emotionally undefined content were used to produce all emotions by ten professional actors (5 female, 5 male). The

This project was funded by AFRL under a subcontract to RADC Inc. under contract FA8750-09-C-0067.

database is recorded in 16 bit, 16 kHz under studio noise conditions. For evaluation, the whole database is randomly divided into two disjoint sets, training and testing (80% for training, 20% for testing).

Table 1. *EMO-DB database.*

Emotion	Session Count	Duration (sec)	Avg. Segment duration (sec)
anger	127	336	2.6
boredom	81	225	2.8
disgust	46	154	3.3
anxiety/fear	69	154	2.2
happiness	71	181	2.5
sadness	62	251	4.0
Neutral	79	187	2.4
Sum	535	1488	2.8

3. Baseline System

The Gaussian Mixture Models (GMM) classifier is a popular method for text independent speaker recognition and has been used for accent identification and EID. We use this approach as our baseline system. Figure 1 shows the block diagram of the baseline GMM training/testing system. The silence removal module sets aside silence in the audio files that are used for training and testing. A parallel gender ID system is used to select emotion sets for each gender (we do not use the gender information from the testing labels in order to mimic real life case). Next, gender dependent GMM are trained for each emotion. While testing, the incoming audio is classified as a particular emotion based on the maximum posterior probability measure over all the GMM candidates. Throughout this paper, we use 64 mixtures for all GMM to provide a fair basis for comparison of different feature extraction front-end.

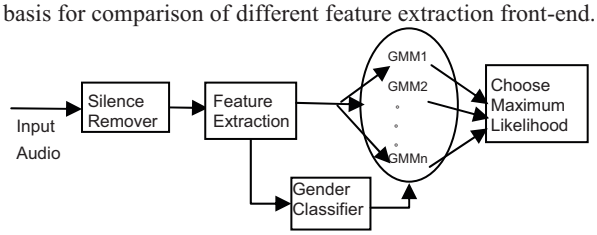

Figure 1: *Baseline GMM based EID system.*

4. Feature Front-end

Psychologists and speech scientists have conducted extensive research in identifying speech emotion differences. Features such as pitch and intensity have been explored which were found to be useful for emotion recognition [7, 8, 15]. Here, we consider a combination of spectral and temporal information for EID. We believe that our features are capable of capturing more salient characteristics in EID tasks.

4.1. Spectral Based Features

4.1.1. MFCC

Many researchers have used spectral based features such as MFCC for the purpose of EID. In our study, an analysis window of 25msec duration is used, with 10msec skipping rate. We use traditional 39-dimensional feature vector consisting of 12-dim MFCC, 12-dim ΔMFCC, 12-dim ΔΔMFCC, 1-dim Energy, 1-dim ΔEnergy, and 1-dim ΔΔEnergy. This feature is used in the baseline system.

4.1.2. PMVDR

Our previous research [16] showed that perceptual Minimum Variance Distortionless Response (PMVDR) feature extraction is better able to model the upper spectral envelope at the perceptually important harmonics, which may include important emotion clues. PMVDR cepstral coefficients provide improved accuracy over traditional MFCC parameters by better tracking the upper envelope of the speech spectrum. Unlike MFCC parameters, PMVDRs do not require an explicit filterbank analysis of the speech signal. We have found this new feature representation provides not only robustness against noise in speech recognition, but also higher accuracy in clean speech tasks. Here, we propose to test this feature in the context of EID. A block diagram of the PMVDR feature extraction [16] is shown in Figure 2.

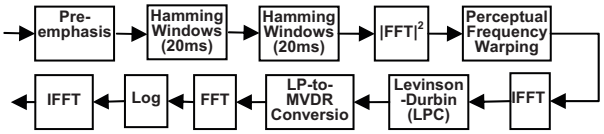

Figure 2: *PMVDR feature extraction process.*

It has been shown that implementing the perceptual scales through the use of a first order all-pass system is feasible [18, 19]. In fact, both Mel and Bark scales are determined by changing the only parameter, α, of the system [18]. The fiter, H(z), and the phase response, β(ω), are given as

$$H(z) = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, |\alpha| < 1 \quad (1)$$

$$\beta(\omega) = \tan^{-1} \frac{(1-\alpha^2)\sin\omega}{(1+\alpha^2)\cos\omega - 2\alpha} \quad (2)$$

where ω represents the linear frequency while β(ω) represents the warped frequency. α controls the degree of warping. For 16 kHz sampled signals, α = 0.42 and 0.55 approximate the Mel and Bark scales, respectively.

Utilizing direct warping on the FFT power spectrum by removing the filterbank processing step leads to the preservation of almost all the information in the short-term speech spectrum. We can now summarize the remainder of the proposed PMVDR algorithm as follows:

1) Obtain the perceptually warped FFT power spectrum,
2) Compute the "perceptual autocorrelations" by utilizing the IFFT on the warped power spectrum,
3) Perform a i^{th} order LP analysis via Levinson-Durbin recursion using perceptual autocorrelation lags,
4) Calculate the i^{th} order MVDR spectrum using Eq.(2) from the LP coefficients [17],
5) Obtain the final cepstrum coefficients using the straightforward FFT-based approach.

Finally, we use 39-dimensional PMVDR features and each feature vector contains 12 statics, deltas and delta-deltas along with energy, delta and delta-delta energy. We used the same windowing and frame skipping as in MFCC before further processing. Cepstral mean normalization was also utilized on the final feature vectors. Since PMVDR remove the filterbank processing, we can avoid the demanding computation and noise sensitivity incurred filterbank processing. This is crucial to realistic emotion identification system.

4.2. Temporal Based Features

4.2.1. Pitch and Intensity

Pitch and intensity are tradition supra-segmental features used in EID since they can capture temporal characteristics, which is one of important characteristics of emotion expression.

4.2.2. SDC

The goal to include shifted delta cepstrum (SDC) in the context of emotion recognition is to incorporate additional temporal information into the feature vector. The SDC is in fact k blocks delta cepstrum coefficients [13]. Suppose the basic set of cepstrum coefficients, $\{c_j(t), j = 1, 2, \ldots, N-1\}$, is available (which can be MFCC or PMVDR statics in this study) at frame t, where j is dimension index and N the number of cepstrum coefficients. The SDC feature can be expressed as

$$s_{(iN+j)}(t) = c_j(t + iP + d) - c_j(t + iP - d), \quad i = 0,1,\ldots,k-1 \tag{3}$$

where d is the time difference between frames for spectra computation, P is the time shift between each block, and k is the total number of blocks. The SDC coefficients can be concatenated with the basic cepstrum coefficients. Thus, we can obtain the feature vector as $\{c_j(t), j=0, 1, \ldots, N-1; s_{(iN+j)}(t), j=0, 1, \ldots, N-1, i=0, 1, \ldots, k-1\}$, which is the SDC version of features. The computation is illustrated in Figure 3.

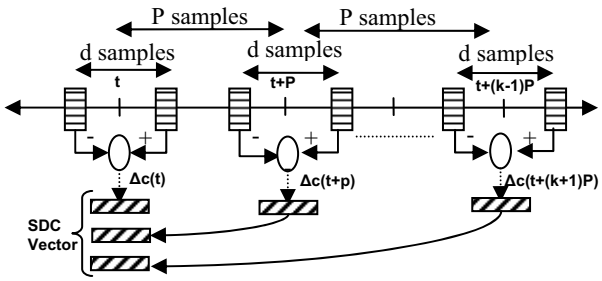

Figure 3. *Computation of the SDC feature vector at frame t for parameters N-d-P-k. The horizontal hatched box means the basic cepstrum coefficients, diagonal hatched box delta feature vector.*

The popular parameter configuration of SDC is N-d-P-k in LID is 7-1-3-7. In our EID task, we fix the optimal configuration Dim-1-3-3 through hill-climbing algorithm, where Dim is the dimension of basic cepstrum coefficients and equals to 11 for both MFCC and PMVDR.

4.3. Multiple Stream Feature for Fusion and SDC

To fully exploit the acoustic information for EID in the feature extraction front-end, we can further combine the diverse feature streams. We can fuse MFCC and PMVDR, respectively, with intensity and pitch to get the MIP and PIP sets. In addition, we can also derive the SDC versions of MFCC, PMVDR, MIP, and PIP by following Sec. 4.2.2. The goal is to compare the contribution of multiple feature streams with SDC to EID tasks. For MIP-SDC, we first derive the MFCC-SDC, and then we combine it with pitch and intensity features. This also applies to PIP-SDC.

5. Experimental Results

5.1. Results

Now we consider an evaluation of the effectiveness of the proposed feature extraction schemes. All experiments are based on the same experiment setup as in Sec. 3, and MFCC is the feature for the baseline system. The results are displayed in Tab. 2 and Fig. 4. For short, MIP stands for MFCC+ Intensity+ Pitch; PIP stands for PMVDR+Intensity+Pitch.

Table 2. *Emotion identification accuracy (%) with Different feature front-end scheme*

	Without SDC	With SDC
MFCC	**71.6**	80.1
MIP	73.8	73.8
PMVDR	73.8	**81.9**
PIP	78.3	79.2

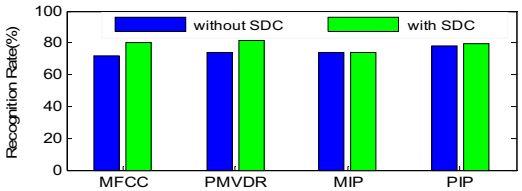

Figure 4. *Bar graph for eight feature extraction front-end performance (2 basic features: MFCC, PMVDR; 2 multi-stream features: MIP, PIP and their SDC version).*

Now we also want to compare the robustness of the different feature extraction schemes in noisy condition. We introduce 2 kinds of additive Gaussian white noise (10dB and 0dB) to the EMO-DB database. We summarize the results in Fig. 5 and Fig. 6 for all the eight feature extraction schemes.

5.2. Analysis

From Table 2 and Figure 4 we can see that both PMVDR and its combined version, PIP, perform better than their corresponding competitors: MFCC and MIP, respectively. PIP is the best one of the four schemes without SDC. When the SDC is applied to the previous four feature vectors, greater improvements are achieved in basic features, MFCC and PMVDR, than in their fused versions. SDC has less contribution to the MIP-SDC and PIP-SDC. Although both SDC and feature combination schemes supposed to provide more temporal cues to help identification tasks, the union of SDC and feature stream fusion are inferior to the SDC alone. This result may be explained by the redundant information from two different temporal features may need to be well coordinated and synchronized before using them together.

In the noise robustness experiment results (Fig. 5), we see PMVDR well outperforms MFCC. Another important finding is that the use of SDC increases the performance in noisy condition in the four feature extraction schemes (PMVDR, MFCC, PIP, and MIP). This can be explained by the cepstrum coefficients subtraction operation in (3). Note, the multi-stream fused feature shows worse robustness than uni-stream features. This result may be explained by the temporal sensitivity introduced by intensity and pitch features. One interesting result is that PMVDR seems to be excited about noise and performs better in 10dB condition than in clean condition, which deserves further exploration.

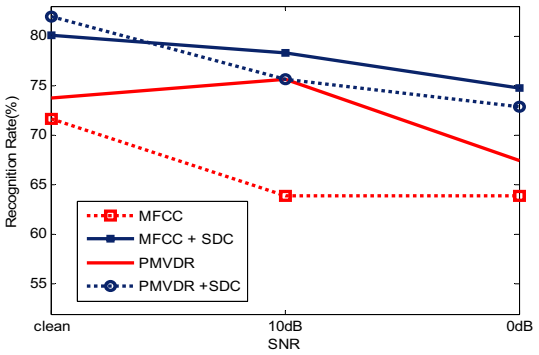

Figure 5. *Robustness performance of the different basic feature extraction front-end.*

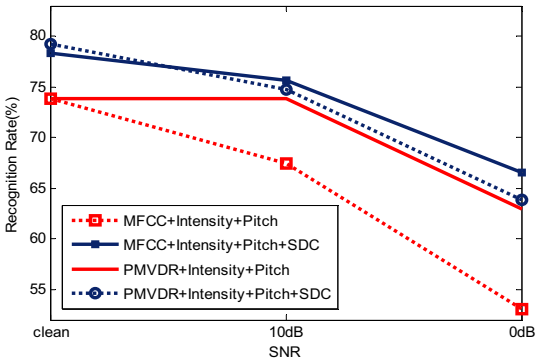

Figure 6. *Robustness performance of the different multi-stream feature fusion front-end.*

6. Conclusions

To perform EID on short utterances less than 3 seconds is a challenge. An effective feature front-end is an important component of a good identification system. The proposed PMVDR-SDC feature extraction, which has not been used in the context of EID, outperforms the baseline system with an absolute gain of 10.3%. When the proposed PMVDR+SDC based EID system's score is linearly combined with that of the baseline (best recall rate: 89.5%) in [8], with the weighting parameter 0.5 for every individual system, we can get an identification accuracy of 91.9%. This partly proved the efficiency of the proposed feature extraction from another perspective.

The improvement may be attributed to PMVDR in that it can better approximate the perceptual scales and less sensitive to speaker variation. Another advantage introduced by PMVDR is its robustness to noise. This result agrees well with previous work reported in the context of automatic speech recognition [16].

SDC is another efficient way to integrate longer temporal information to compensate the basic cepstral coefficient features. We show that the use of this feature enhances the system performance both in accuracy and robustness. SDC works much better in the basic feature sets than in multiple feature stream combinations and is simpler than delicate feature fusion schemes. These results are concrete steps towards building a real life EID system.

Future exploration on the possibilities of integrating more spectral and temporal information may help to build a more effective emotion recognition system. We will also conduct experiment with spontaneous (non-acted) databases.

7. Acknowledgements

We would like to thank Dr. Sanjay A Patil and Dr. Carlos Busso for their comments during this study.

8. References

[1] J.H.L. Hansen, "Analysis and Compensation of Speech Under Stress and Noise for Environmental Robustness in Speech Recognition," Speech Communication, 20 (2), pp.151-170, 1996.

[2] Wooil Kim and J. H. L. Hansen, "Robust Angry Speech Detection Employing TEO-Based Discriminative Classifier Combination," Interspeech-2009, pp. 2019-2022, Brighton, UK, Sept. 2009.

[3] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion Recognition in Human-Computer Interaction," IEEE Signal Processing Magazine, vol.18, no.1., pp.32-80, 2001.

[4] M. Rahurkar, J.H.L. Hansen, J. Meyerhoff, G. Saviolakis, and M. Koenig, "Frequency Band Analysis for Stress Detection Using a Teager Energy Operator Based Feature," ICSLP-2002, pp.2021- 2024, 2002.

[5] L. Devillers, and L. Vidrascu, "Real-life Emotions Detection with Lexical and Paralinguistic Cues on Human-Human Call Center Dialogs," Interspeech2006, 2006.

[6] V. Sethu, E. Ambikairajah and J. Epps, "Empirical Mode Decomposition based weighted Frequency Feature for Speech-based Emotion Classification," ICASSP2008, pp.5017-5020, 2008.

[7] Oudeyer P-Y. The production and recognition of emotions in speech: features and algorithms, International Journal in Human-Computer Studies, 59(1-2), pp. 157-183, 2003

[8] Florian Eyben, Martin Wollmer, Bjorn Schuller: "openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit", in Proc. 4th International HUMAINE Association Conference on Affective Computing and Intelligent Interaction 2009 (ACII 2009), IEEE, Amsterdam, The Netherlands, 10.-12.09.2009.

[9] Davis, S. B. and Mermelstein, P., "Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences," IEEE Trans. ASSP, Vol 28, pp 357–366, 1980.

[10] Gu, L. and Rose, K., "Perceptual Harmonic Cepstral Coefficients as the Front-end for Speech Recognition", Proc. ICSLP'00

[11] Bou-Ghazale, S. E., Hansen, J. H. L., "A Comparative Study of Traditional and Newly Proposed Features for Recognition of Speech Under Stress," IEEE Trans. Speech & Audio Proc., vol. 8, pp. 429-442, July 2000.

[12] Haizhou Li, etc. "The I4U system in NIST 2008 speaker recognition evaluation," ICASSP, pp.4201-4204, 2009.

[13] B. Bielefeld, "Language identification using shifted delta cepstrum," In 14th Annual Speech Research Symposium, 1994.

[14] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter Sendlmeier und Benjamin Weiss, "A Database of German Emotional Speech," Proc. Interspeech 2005, Lissabon, Portugal

[15] B. Schuller, S. Reiter, and G. Rigoll. Evolutionary Feature Generation in Speech Emotion Recognition. In: Proc. Int. Conf. on Multimedia & Expo ICME 2006, Toronto, Canada, pp. 5–8. IEEE, 2006. 09.-12.07.2006.

[16] Umit H. Yapanela, J.H.L. Hansen. "A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition". Speech Communication 50 (2008) 142–152.

[17] Murthi, M. N. and Rao, B. D., "All-pole modeling of speech based on the minimum variance distortionless response spectrum," IEEE Trans. Speech & Audio Proc., May 2000.

[18] Tokuda, K., Masuko, T., Kobayashi, T., Imai, S., " Mel-generalized Cepstral Analysis-A Unified Approach to Speech Spectral Estimation", Proc.ICSLP'94

[19] Smith, J. O. and Abel, J. S., "Bark and ERB Bilinear Transforms", IEEETrans. Speech & Audio Proc., Nov. 1999