INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Front-end Feature Compensation and Denoising for Noise Robust Speech Emotion Recognition

Rupayan Chakraborty, Ashish Panda, Meghna Pandharipande, Sonal Joshi, Sunil Kumar Kopparapu

TCS Research and Innovation, Thane West, Mumbai, INDIA
{rupayan.chakraborty,ashish.panda,meghna.pandharipande,
sonals.joshi,sunilkumar.kopparapu}@tcs.com

Abstract

Front-end processing is one of the ways to impart noise robustness to speech emotion recognition systems in mismatched scenarios. Here, we implement and compare different front-end robustness techniques for their efficacy in speech emotion recognition. First, we use a feature compensation technique based on the Vector Taylor Series (VTS) expansion of noisy Mel-Frequency Cepstral Coefficents (MFCCs). Next, we improve upon the feature compensation technique by using the VTS expansion with auditory masking formulation. We have also looked into the applicability of 10^{th}-root compression in MFCC computation. Further, a Time Delay Neural Network based Denoising Autoencoder (TDNN-DAE) is implemented to estimate the clean MFCCs from the noisy MFCCs. These techniques have not been investigated yet for their suitability to robust speech emotion recognition task. The performance of these front-end techniques are compared with the Non-Negative Matrix Factorization (NMF) based front-end. Relying on extensive experiments done on two standard databases (EmoDB and IEMOCAP), contaminated with 5 types of noise, we show that these techniques provide significant performance gain in emotion recognition task. We also show that along with front-end compensation, applying feature selection to non-MFCC high-level descriptors results in better performance.

Index Terms: Speech emotion recognition, noise robustness, denoising autoencoder, feature compensation, auditory masking

1. Introduction

Environmental noise corrupts acoustic cues and therefore degrades the performance of speech emotion recognition systems. [1–4]. In literature, noise in emotion recognition has been handled by techniques such as enhancing speech signals, eliminating noise, adapting models, compensating features and deriving robust set of acoustic features. Histogram equalization to reduce the difference between feature vectors in clean and noisy conditions have been proposed in [5]. In [1], Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and their combinations have been explored for different types of noises. In [6], Fisher rate to PCA for dimension reduction has been used with an ANN classifier. Authors in [2], extracted $4k$ acoustic features, reduced them by fast Information-Gain-Ratio (IGR) filter-selection according to different noise type and finally classified using a SVM classifier. In [7], effect of word- or turn-based features with different noise addition and microphone position has been explored. Spectral subtraction along with masking has been used to enhance the signal in [8] and its effect was studied on white noise contaminated speech. Discarding non-speech or noise dominated frames by using a simple voice activity detector (VAD) has been shown to be effective in [3, 4].

In this paper, we implement and compare a series of different front-end robustness techniques for speech emotion recognition task. First, we implement a VTS based front-end technique to estimate robust MFCC features. The VTS technique relies on a Gaussian Mixture Model (GMM) trained on clean speech and a vector Taylor series expansion of the noisy model means to estimate the GMM of noise corrupted speech. Using the clean and the noise corrupted GMMs, a robust MFCC feature is estimated from a noisy MFCC feature. The estimated robust MFCC feature can be used for speech emotion recognition to obtain better performance. Next, we look into the auditory masking formulation of the VTS (VTS-AM) and show that it improves upon the performance of the VTS. We also compare the performance of the MFCC features with log-compression and 10^{th}-root compression. Second, we train a TDNN-DAE to estimate the clean counterpart of a noisy MFCC vector. Further, we also look into the applicability of feature selection technique for non-MFCC high level descriptors and show that they provide a small but consistent performance gain on top of what is obtained by the other front-end techniques. Moreover, to find the effectiveness of our proposed system in more realistic conditions, we choose to experiment in mismatched scenarios, i.e *clean*-training and *noisy*-testing. Comparison of these techniques with the ones proposed in [3, 4, 9], shows just how effective these techniques can be in imparting robustness to the emotion recognition systems. Although, different feature compensation and model adaptation techniques have been used in noisy speech emotion recognition task (e.g. [5, 8, 10]), the methods used in this paper, such as VTS, VTS-AM and root compression have never been investigated for speech emotion recognition task. Since VTS and VTS-AM have been shown to be effective in speech recognition tasks [11], it is interesting to see how they perform for emotion recognition task. Experiments with two different databases and with 5 different types of noises, each at 5 different levels, show that the proposed systems performed remarkably well. It should be mentioned that we have not investigated the Lombard effect in this work. However, the proposed methods should work to counter the degradation caused by additive noise.

The rest of the paper is organized as follows. Section 2 presents emotion recognition system, along with our proposed feature compensation and selection technique. In section 3, we explain the experimental setup, databases, results, and analysis. Conclusion is given in section 4.

2. Feature compensation and denoising for noisy speech emotion recognition

Here we propose the emotion recognition system for noisy speech (as depicted in Figure 1) that consists of a feature extraction at the front end, followed by a conventional classifier. Fea-

http://dx.doi.org/10.21437/Interspeech.2019-2243

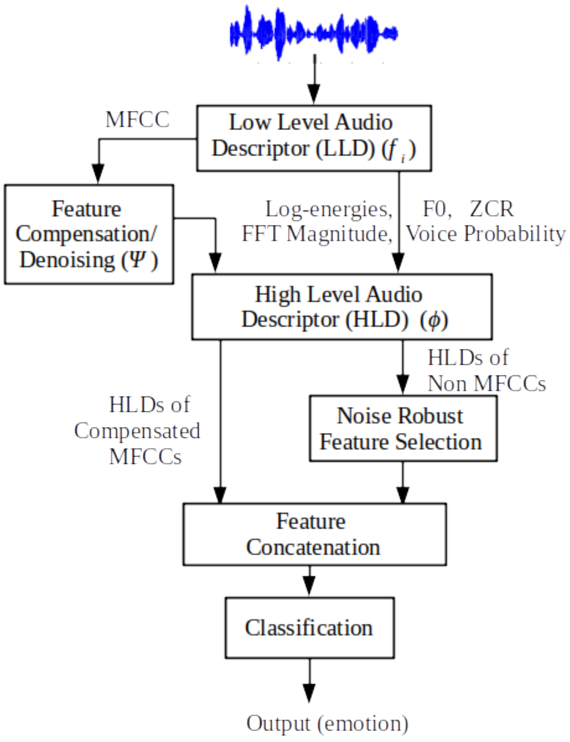
MFCC
Low Level Audio Descriptor (LLD) (f_i)
Feature Compensation/ Denoising (Ψ)
Log-energies, FFT Magnitude,
F0, ZCR Voice Probability
High Level Audio Descriptor (HLD) (ϕ)
HLDs of Non MFCCs
HLDs of Compensated MFCCs
Noise Robust Feature Selection
Feature Concatenation
Classification
Output (emotion)

Figure 1: *Front-end feature compensation (or denoising) and selection for noisy SER.*

ture compensation and selection are proposed within the feature extraction module to deal with the noisy speech.

Let us denote s be the noisy speech. Overall feature extraction process is defined by,

$$\zeta = \phi_j\{f_i(s)\}_{j=1,i=1}^{J,N} \tag{1}$$

where f_i and ϕ_j are the low-level and the high-level feature extraction functionals. N and J are total number of low and high level functionals respectively. In this work, the main objective is to compensate and select noise robust features at different stages of the feature extraction module. Therefore, we, essentially, rewrite the Equation 1 as,

$$\zeta = \{\phi_j\{\psi(f_1(s))\}, \hat{\phi}_j\{f_i(s)\}\}_{j=1,i=2}^{J,N} \tag{2}$$

where ψ is the compensation functional which give compensated version of MFCC features (f_1). f_i are the non-MFCC feature extraction functionals. ϕ and $\hat{\phi}$ are the high-level and selected high-level features, respectively.

2.1. VTS and VTS-AM

Traditional assumption of noise corruption model is that the speech and noise are additive in the spectral magnitude domain. While compensating through VTS expansion [11, 12], non-linear function in cepstral domain can be represented as,

$$y^s = x^s + h^s + Clog(1 + exp(C^{-1}(n^s - x^s - h^s))) \tag{3}$$

where C and C^{-1} are the Discrete Cosine Transformation (DCT) matrix and it's inverse, respectively. On the other hand, $\boldsymbol{y}$, $\boldsymbol{x}$, $\boldsymbol{h}$ and $\boldsymbol{n}$ are the MFCC domain distorted speech, clean speech, channel factor, and additive noise parameters. But, according to psychoacoustic corruption model [13], only the portion of noise which is above the masking threshold of clean speech is added to the speech. The reason being that the masked portion of the noise is not audible and hence plays little role in the distorted speech. The psychoacoustic corruption function (as described in [14]) is used to modify the Equation 3 by incorporating the auditory masking criteria (i.e VTS-AM), as follows:

$$y^s = x^s + h^s + w^s + Clog(1 + exp(C^{-1}(n^s - x^s - h^s - w^s))) \tag{4}$$

where $\boldsymbol{w}$ is a scaling factor, which depends on the masking threshold of the clean speech. Compensated model parameters can be computed by following similar methods as described in [12, 15]. The modified Taylor series component G which is the Jacobian of the mismatch function is defined as:

$$G = C.diag\left(\frac{1}{1 + exp(C^{-1}(\boldsymbol{\mu_n} - \boldsymbol{\mu_x} - \boldsymbol{w} - \boldsymbol{h}))}\right).C^{-1} \tag{5}$$

where the component G is derived using only the static portion of model mean and noise mean. Next, we compensate the model mean and variance as follows:

$$\boldsymbol{\mu}_y = \boldsymbol{\mu}_x + \boldsymbol{h} + \boldsymbol{w} + Clog(1 + exp(C^{-1}(\boldsymbol{\mu_n} - \boldsymbol{\mu}_x - \boldsymbol{w} - \boldsymbol{h}))) \tag{6}$$

and

$$\Sigma_y \approx G\Sigma_x G^T + (I - G)\Sigma_n(I - G)^T \tag{7}$$

where I and T are the identity matrix and transpose respectively. $\boldsymbol{\mu}_y$ and Σ_y are the compensated mean and variance. In this approach, a GMM is trained on the clean speech and is denoted as $\lambda_x = \{\boldsymbol{\mu}_x, \boldsymbol{\sigma}_x, \boldsymbol{w}\}$. Next, the GMM parameters (mean and variance) are compensated according to the method described in [11]. Let the compensated model be denoted as $\lambda_y = \{\boldsymbol{\mu}_y, \boldsymbol{\sigma}_y, \boldsymbol{w}\}$. The pseudo-clean features $\boldsymbol{x}_{MMSE}$ are estimated from the noisy observations as [16]:

$$\boldsymbol{x}_{MMSE} = \boldsymbol{o} - \sum_{m=0}^{M-1} p(\boldsymbol{o}|\lambda_{ym})(\boldsymbol{\mu}_{ym} - \boldsymbol{\mu}_{xm}) \tag{8}$$

where $\boldsymbol{o}$ is the noisy speech features. $p(\boldsymbol{o}|\lambda_{ym})$ is the posterior probability for the m^{th} Gaussian mixture component of the noise compensated GMM against the observation $\boldsymbol{o}$. $\boldsymbol{\mu}_{ym}$ is the m^{th} component of the noise compensated GMM and $\boldsymbol{\mu}_{xm}$ is the m^{th} component of the clean GMM.

2.1.1. Log and root compression

In MFCC feature computation, usually logarithm is used on top of the mel-filterbank outputs. The purpose of applying logarithm is to reduce the dynamic range of the feature and also to make data less sensitive towards variability [11, 17]. However, root compression can also be used to achieve the same goal. The benefit of logarithmic is that channel effect can be discarded through cepstral mean and variance normalization (CMVN), which is not possible for root compression. However, since in our emotion recognition framework, we do not use CMVN, using root-compression might provide better peformance.

2.2. Denoising autoencoder

For denoising the MFCC features, we used denoising autoencoders (DAE) with a TDNN architecture [18]. TDNN architecture has a narrow context for initial layers and wider context for deeper layers, which helps the TDNN to learn the features and the temporal relationships in a long term temporal context and therefore the architecture is expected to perform well as a DAE [19]. We have followed the TDNN-DAE network architecture used in [20]. Further details of the TDNN-DAE are given in the Section 3.

2.3. Feature selection and high-level audio descriptors

We have applied a robust feature selection algorithm for non-MFCC high-level descriptors (HLDs) that selects the relevant features in diverse noise conditions. Information Gain Ratio based feature selection (IGR-FS), where highly relevant attributes are found by their entropy, has been used for this purpose. The ranking of attributes obtained by IGR-FS is independent of the classifier [2]. We also tried Correlation based feature selection (CFS) [21]. But, better performances were observed for IGR-FS than the CFS-based feature selection, and the former one is computationally faster as well. It is to be noted that IGR-FS on MFCC features did not improve the performance and hence we restricted it to non-MFCC HLDs. High-level descriptors are computed on top of all the low-level audio descriptors, which consist of compensated (or denoised) MFCCs and non-MFCC descriptors.

3. Experiments

3.1. Database and experimental set-up

We experimented with 2 standard emotional databases, namely (1) Berlin emotional database (EmoDB) [22, 23] and (2) Interactive emotional dyadic motion capture database (IEMOCAP) [24, 25], which were contaminated by noise to test our proposed techniques. EmoDB consists of 535 utterances, where 10 professional actors participated to act for 7 emotions. IEMOCAP is having 12 hours of audiovisual data, based on improvised and scripted interactions between 5 pairs of male-female participants. We have taken 5 types of noise (Voice babble,Factory noise, HF radio channel, F-16 fighter jets, and Volvo 340) from Noisex-92 database to corrupt the clean speech [26]. FANT toolkit is used for contamination of noise to clean speech at 5 SNR levels (0dB ,5dB, 10dB, 15dB, and 20dB) [27].

In all our experiments, we extracted 6 LLDs (frame-length of 20 ms and frame-shift of 10 ms), namely, log-energies, voice probability, frequency-band energies, F0, ZCR, and MFCC. On top of all LLDs, we took 39 statistical functionals (up to fourth order) to extract HLDs. We have used openSMILE toolkit for extracting acoustic features other than the MFCCs [28]. 23 dimensional MFCC features (with Δ and $\Delta\Delta$) were extracted for feature compensation using Kaldi speech recognition toolkit [29]. Noisy features are compensated using VTS and VTS-AM, and then transformed from the MFCC domain to mel-filter bank domain. Next, we apply log and 10^{th}-root compression on the VTS compensated features. However, for both compensated and non-compensated MFCC features, we extract the HLDs.

For denoising, we used Kaldi based TDNN-DAE architecture, which has 4 hidden layers and each hidden layer consists of 1024 ReLU activation nodes [20]. Contexts for the DAE network with four hidden layers is organized as (-2,-1,0,1,2) (-1,2) (-3,3) (-7,2) (0) and the input temporal context to [-13,9]. We train the TDNN-DAE using noisy speech utterances (5 types of noise at 5 SNR levels) and their corresponding clean speech utterances. During test, the noisy test features are passed through the trained DAE to get the denoised output features.

3.2. Results and analysis

All our experiments have been conducted on clean-training and noisy-testing, which is a mismatched scenario. For all our experimentation, we followed 5 cross-validation (CV) setup by splitting up the dataset into 5 sets (80%-20% for train-test), and following leave one out for (noisy) testing in each validation. Noise was added only to the test set (20% of the data) for each validation. We trained a new TDNN-DAE for each validation, so that test samples are never seen during the training. Emotion recognition accuracies (RA in %) for the two datasets and for different techniques are tabulated in Table 1. We also experimented without selecting non-MFCC features concatenated with MFCC compensated (or denoised) features, but feature selection (i.e using IGR-FS) always performed better. Due to space constraint we could not include all the results in this table, but major results are covered.

It can be noted from Table 1 that the trend is similar in both Emo-DB and IEMOCAP. A small gain is consistently obtained by using NMF based denoising method. The gain is further improved by using the energy based VAD. This is probably because the non-speech/silence frames do not contain significant emotional information and are easily dominated by noise. The energy based VAD helps in improving performance by discarding these spurious non-speech frames. It should be noted here that this energy based VAD outperformed the openSMILE RNN-based VAD in our earlier work [3, 4]

The log-compressed MFCC features (using VTS) and selected non-MFCC provide better performance gain than the NMF+VAD combination. This indicates superior estimation of robust features by VTS. Although we have not reported here the performance of VTS without feature selection of non-MFCC, we have observed that it still outperforms the NMF+VAD combination. Feature selection improves the VTS performance by a small but consistent margin. It can be seen that the VTS-AM formulation outperforms the traditional VTS formulation. This shows the efficacy of acoustic masking in estimating robust features and it is in-line with what has been reported in the ASR domain.

The 10^{th}-root-compressed MFCC features along with VTS-AM provide the best performance of all the unsupervised algorithms. In this case also, the proposed feature selection technique provides a small but consistent gain over VTS-AM alone. Table 1 reports the results from 10^{th}-root-compressed MFCC features along with VTS-AM and the selected non-MFCC features. The 10^{th}-root-compression has resulted in a significant performance gain over the log-compression and it suggests that 10^{th}-root-compression should be considered in cases where cepstral mean and variance normalization is not required.

The supervised TDNN based DAE provides better performance than any of the unsupervised method described in this paper. This is probably because the TDNN-DAE, as has been used here, operates in seen noise conditions, i.e, the noise encountered in the test utterances are the same type that has been used to train the DAE, although the speech utterances are different. As has been shown in [30, 31], in case of unseen noise conditions the TDNN-DAE may not perform as nicely as has been seen here. Overall, it can be inferred from Table 1 that the feature denoising and selection methods described in this paper make the emotion recognition system robust against noisy utterances.

Table 1: *Categorical emotion RA (in %) (5 types of noise with 5 SNR levels) for different systems: No Comp: No compensation, $\zeta 1$: log-MFCC + VTS+$\hat{\phi}$, $\zeta 2$: log-MFCC + VTS-AM+$\hat{\phi}$, $\zeta 3$: 10^{th} root-MFCC + VTS-AM+$\hat{\phi}$, $\zeta 4$: TDNN-DAE+$\hat{\phi}$, $\hat{\phi}$: Feature selection on non-MFCC features*

		EmoDB							**IEMOCAP**						
Noise type	**SNR**	**No Comp.**	**NMF**	**NMF + VAD**	$\zeta 1$	$\zeta 2$	$\zeta 3$	$\zeta 4$	**No Comp.**	**NMF**	**NMF + VAD**	$\zeta 1$	$\zeta 2$	$\zeta 3$	$\zeta 4$
Babble	0dB	20.9	21.54	23.13	23.12	35.45	37.45	**37.86**	20.18	21.11	23.42	25.97	23.32	27.91	**30.42**
	5dB	22.3	23.09	25.26	37.01	44.54	46.72	**47.92**	22.07	22.21	20.13	28.76	29.43	30.41	**32.53**
	10dB	24.54	25.81	27.12	47.89	52.71	57.27	**67.54**	30.51	26.32	27.41	29.01	31.34	35.66	**39.67**
	15dB	28.18	29.09	33.42	56.63	66.36	69.19	**69.94**	32.46	28.85	28.18	30.66	34.21	37.01	**41.06**
	20dB	35.45	37.9	42.13	62.67	70.9	73.68	**79.71**	32.51	33.12	33.52	33.16	35.33	38.16	**46.88**
F16	0dB	13.63	14.98	18.34	17.12	27.27	32.72	**42.57**	11.68	17.62	16.21	20.92	21.52	23.92	**25.48**
	5dB	18.18	19.09	19.32	21.76	30.16	37.27	**40.85**	14.28	19.62	19.23	21.26	22.43	27.76	**31.67**
	10dB	20.9	21.11	23.54	34.31	42.72	56.36	**57.23**	16.23	21.32	20.2	25.61	29.76	31.46	**33.13**
	15dB	25.45	27.32	30.13	50.98	50.9	58.9	**65.53**	27.27	28.09	29.09	29.92	32.08	34.41	**40.14**
	20dB	34.54	35.18	39.13	57.34	59	65.82	**67.34**	27.92	29.11	30.81	30.73	34.52	37.66	**42.03**
Factory	0dB	12.72	15.11	17.11	29	29.09	36.36	**39.09**	22.72	22.12	23.11	24.18	26.98	29.31	**30.23**
	5dB	17.27	19.08	22.35	38.09	44.54	56.36	**58.13**	24.67	25.22	26.01	26.48	28.42	32.52	**33.78**
	10dB	21.81	22.87	25.33	48.12	55.45	57.29	**59.12**	25.32	27.09	27.91	29.92	32.41	35.76	**38.18**
	15dB	24.54	27.32	29.42	59.96	58.79	62.72	**64.48**	30.51	30.76	31.77	32.67	36.12	38.92	**41.54**
	20dB	32.72	39.51	42.84	64.5	63.63	68.59	**69.96**	31.81	31.22	33.06	34.92	40.08	43.92	**47.53**
HFchannel	0dB	20	23.12	23.32	18.23	40.9	54.12	**54.62**	22.72	23.23	23.87	24.07	30.09	32.57	**34.77**
	5dB	21.81	25.33	25.34	28.21	45.45	58.31	**59.28**	30.51	31.21	26.66	27.32	32.31	35.12	**39.13**
	10dB	24.54	27.65	28.32	42.23	56.36	60.75	**64.49**	29.22	30.87	30.13	32.26	39.21	47.41	**51.23**
	15dB	34.54	37.52	39.63	54.44	65.45	68.43	**69.71**	33.12	37.12	36.12	39.36	45.12	52.12	**53.08**
	20dB	48.18	52.21	54.53	62.31	70.9	73.42	**76.36**	34.41	38.21	39.87	45.62	52.32	58.01	**60.32**
Volvo	0dB	15.54	20.22	21.42	28.09	48.18	56.9	**60.17**	16.23	20.66	26.88	33.67	34.55	36.87	**39.63**
	5dB	23.63	27.65	29.34	41.32	53.63	60.36	**63.41**	18.23	27.13	29.81	35.21	36.66	39.41	**41.68**
	10dB	37.27	36.13	41.21	55.8	61.81	68.18	**70.13**	23.37	30.42	32.67	35.74	37.89	42.94	**44.93**
	15dB	57.27	59.15	62.23	70.32	72.72	74.87	**76.58**	28.57	35.51	37.76	41.18	42.34	49.48	**52.24**
	20dB	62.72	67.22	67.23	74.45	74.54	76.21	**80.43**	29.87	38.21	40.32	42.43	44.51	53.83	**59.16**

4. Conclusion

In this paper, we propose front-end feature denoising and compensation techniques for noise robustness in speech emotion recognition task. TDNN-DAE based feature denoising is found to be performing the best in seen noise condition. VTS based feature compensation with psychoacoustic masking has also been proved to be beneficial for front-end processing, but not performing like TDNN-DAE in seen noise conditions. While computing MFCC features, 10^{th}-root compression found to gel well with VTS-AM in comparison with the log compression. It is to be noted here that the TDNN-DAE is operating in seen noise conditions (i.e the type of noise encountered in the test utterances has been used to train the DAE, although not the utterance itself). For unseen noise conditions, the TDNN-DAE might not perform as VTS or VTS-AM, like in speech recognition task. We intend to tackle this in our future studies. It can also be observed that applying feature selection technique to non-MFCC high-level descriptors on top of the proposed compensation techniques provides a small but consistent gain in performance. Moreover, the proposed methods outperform previously used NMF-based enhancement or even the NMF-VAD by a significant margin, clearly indicates its efficacy.

5. References

[1] M. You, C. Chen, J. Bu, J. Liu, and J. Tao, “Emotion recognition from speech signal combining PCA and LDA,” 2006.

[2] B. Schuller, D. Arsi, F. Wallhoff, and G. Rigoll, “Emotion recognition in the noise applying large acoustic feature sets,” *Speech Prosody*, 2006.

[3] M. Pandharipande, R. Chakraborty, A. Panda, and S. Kopparapu, “An unsupervised frame selection technique for robust emotion recognition in noisy speech,” in *Proc. EUSIPCO*, 2018.

[4] M. Pandharipande, R. Chakraborty, A. Panda, and S. K. Kopparapu, “Robust front-end processing for emotion recognition in noisy speech,” in *11th International Symposium on Chinese Spoken Language Processing, ISCSLP 2018, Taipei City, Taiwan, November 26-29, 2018*, 2018, pp. 324–328.

[5] L. Juszkiewicz, “Improving noise robustness of speech emotion recognition system,” *Intelligent Distributed Computing VII*, 2014.

[6] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, “Speech emotion recognition: Features and classification models,” *Digital Signal Processing*, 2012.

[7] B. Schuller, D. Seppi, A. Batliner, A. K. Maier, and S. Steidl, “Towards more reality in the recognition of emotional speech,” *IEEE ICASSP*, 2007.

[8] C. Huang, G. Chen, H. Yu, Y. Bao, and L. Zhao, “Speech emotion recognition under white noise,” *Archives of Acoustics*, 2013.

[9] F. Weninger, B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognition of nonprototypical emotions in reverberated and noisy speech by nonnegative matrix factorization," *EURASIP Journal on Advances in Signal Processing*, 2011.

[10] J. Pohjalainen, F. Fabien Ringeval, Z. Zhang, and B. Schuller, "Spectral and cepstral audio noise reduction techniques in speech emotion recognition," *ACM on Multimedia Conference*, 2016.

[11] B. Das and A. Panda, "Robust front-end processing for speech recognition in noisy conditions," *International Conference on Acoustics, Speech and Signal Processing*, 2017.

[12] J. Li, L. Deng, D. Yu, Y. Gong, and A. Acero, "High-performance hmm adaptation with joint compensation of additive and convolutive distortions via vector taylor series," *IEEE Workshop on Automatic Speech Recognition Understanding (ASRU)*, 2007.

[13] A. Panda and T. Srikanthan, "Psychoacoustic model compensation for robust speaker verification in environmental noise," *IEEE Trans. Audio, Speech & Language Processing*, 2012.

[14] A. Panda, "A fast approach to psychoacoustic model compensation for robust speaker recognition in additive noise," *Proc. INTERSPEECH, Germany*, 2015.

[15] A. Alex, D. Li, K. Trausti, and Z. Jerry, "HMM adaptation using vector taylor series for noisy speech recognition," *Proc. Int. Conf. on Spoken Language Processing*, 2000.

[16] B. Li and K. C. Sim, "Noise adaptive front-end normalization based on vector taylor series for deep neural networks in robust speech recognition," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[17] S. Ravindran, D. V. Anderson, and M. Slaney, "Improving the noise-robustness of mel-frequency cepstral coefficients for speech processing," *SAPA@INTERSPEECH*, 2006.

[18] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," in *Readings in speech recognition*. Elsevier, 1990, pp. 393–404.

[19] C.-T. Do and Y. Stylianou, "Improved automatic speech recognition using subband temporal envelope features and time-delay neural network denoising autoencoder," *Proc. INTERSPEECH*, pp. 3832–3836, 2017.

[20] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. INTERSPEECH*, 2015.

[21] M. A. Hall, "Correlation-based feature subset selection for machine learning," Ph.D. dissertation, University of Waikato, Hamilton, New Zealand, 1998.

[22] "EmoDB- Berlin Database of Emotional Speech," *http://www.emodb.bilderbar.info/*.

[23] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," *Proc. INTERSPEECH*, 2005.

[24] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, 2008.

[25] "IEMOCAP- Interactive Emotional Dyadic Motion Capture Database," *http://sail.usc.edu/iemocap/*.

[26] "NOISEX-92 database," *http://spib.rice.edu/spib/select_noise.html*.

[27] "FaNT- Filtering and Noise Adding Tool," *http://dnt.kr.hsnr.de/download/*.

[28] "openSMILE- toolkit," *http://www.audeering.com/research/opensmile*.

[29] D. Povey, A. Ghoshal, G. Boulianne, O. G. L. Burget, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

[30] B. Das and A. Panda, "Integrating denoising autoencoder and vector taylor series with auditory masking for speech recognition in noisy conditions," *Proc. EUSIPCO*, 2018.

[31] S. Joshi, A. Panda, and B. Das, "Enhanced denoising autoencoder for robust speech recognition in unseen noise conditions," in *Proc. ISCSLP*, 2018, pp. 359–363.