Front-End Feature Compensation for Noise Robust Speech Emotion Recognition

Meghna Pandharipande, Rupayan Chakraborty, Ashish Panda, Biswajit Das, Sunil Kumar Kopparapu
TCS Research and Innovation - Mumbai
Yantra Park, Thane, Maharashtra, INDIA, 400601

***Abstract*—Robust feature compensation and selection are important aspects of noisy speech emotion recognition (SER) task, especially in mismatched condition, when the models are trained on clean speech and tested in the noisy scenarios. Here we propose the use of front-end feature compensation techniques based on Vector Taylor Series (VTS) expansion and VTS with auditory masking (VTS-AM) to improve the performance of SER systems. On top of VTS and VTS-AM, we compare the performances of log-compression and root-compression to the mel-filter-bank energies. Further, we demonstrate the benefit of feature selection applied to the non-MFCC high-level descriptors in conjunction with VTS, VTS-AM and root compression. The system performance is compared with popular Non-negative Matrix Factorization (NMF) based enhancement and energy based voice activity detector (VAD) technique, which discards silence or noisy frames in the spoken utterances. To demonstrate the efficacy of our proposed techniques, extensive experiments are conducted on 2 standard datasets (EmoDB and IEMOCAP), contaminated with 5 types of noise (Babble, F-16, Factory, Volvo, and HF-channel) from the Noisex-92 noise database at 5 SNR levels (0dB, 5dB, 10dB, 15dB and 20dB).**

Index Terms: Emotion recognition, Noisy speech, Feature compensation, Auditory masking, Vector Taylor Series

I. Introduction

Emotion recognition in noisy speech faces exponential challenges, mainly because of the corrupted acoustic cues [1]–[4]. Systems restricted to use speech samples recorded in controlled environments cannot be used for realistic speech emotion recognition task.

In literature, to deal with noise in emotional speech, techniques like enhancing speech signals, eliminating noise, adapting models, compensating features and deriving robust set of acoustic features have been explored. For example, histogram equalization to reduce the difference between features vectors in clean and noisy conditions have been proposed in [5]. In [1], [6], authors used Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and their combinations for different types of noises. In [7], Fisher rate to PCA for dimension reduction have been used with an ANN classifier. Authors in [2], extracted $4k$ acoustic features, reduced them further by fast Information-Gain-Ratio (IGR) filter-selection according to different types of noise, and finally classified using a SVM classifier. In [8], authors show how emotion recognition performances are affected by word- or turn-based features with different noise addition and microphone position.

In [9], authors used spectral subtraction along with masking for speech enhancement in white noise contamination. Work in [3], [4], proposed to use a front-end signal processing for discarding noise affected non-speech frames using VAD.

In this paper, we propose a front-end that extracts multiple low-level features, a portion of which is compensated at low-level itself, followed by a high-level (i.e. statistical) feature extraction and then a subsequent noise-robust feature selection. As feature compensation, we propose to use Vector Taylor Series (VTS) expansions and compare its performance with and without psychoacoustic corruption function (VTS-AM). Both VTS and VTS-AM methods are used to compensate a Gaussian mixture model (GMM) trained on clean speech. Then Minimum Mean Square Errors (MMSE) are used to estimate the clean speech features from noisy speech features. Further, we have compared the performance of log or root-compression to the mel-filter bank energies. We also show that applying feature selection on the non-MFCC high level descriptors, along with the VTS and VTS-AM techniques, provides a small but consistent performance gain.

Although, different feature compensation and model adaptation techniques have been used in noisy speech emotion recognition task (e.g. [5], [9], [10]), the methods used in this paper, such as VTS, VTS-AM and root compression have never been used for speech emotion recognition task. While VTS and VTS-AM have been shown to be effective in speech recognition tasks [11], their applicability, so far, has not been shown in speech emotion recognition. 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. We also compare the performance of our proposed techniques with previous works [3], [4], [12], and to make the comparisons fair, we replicate the results of our previously proposed techniques in mismatched scenario. Experiments with two different databases and with 5 different types of noises, each at 5 different levels, show that the proposed systems performed significantly better than the other systems in the literature. 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 II presents emotion recognition system, along with our proposed feature compensation and selection technique. In section III, we explain the experimental setup, databases, results, and analysis. Conclusion is given in section IV.

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

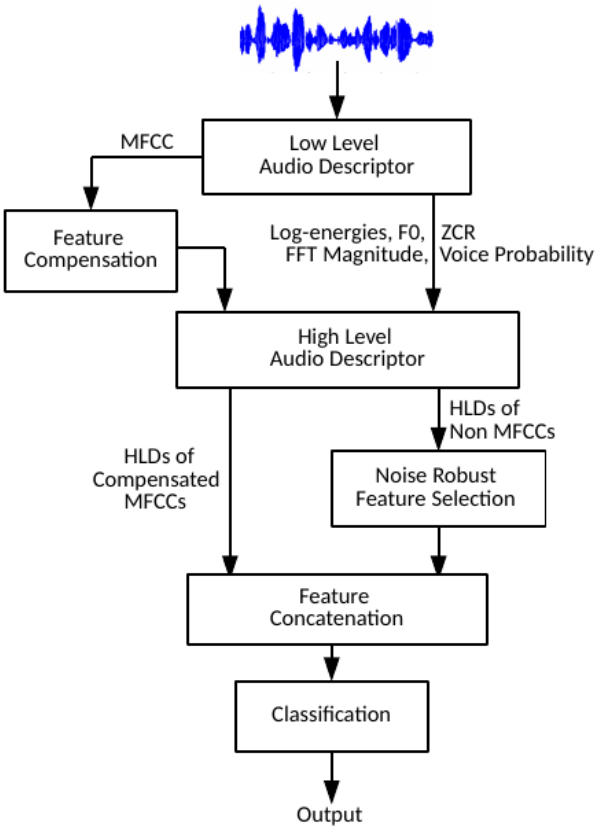

Fig. 1. Front-end feature compensation and selection modules for noisy SER.

II. Feature compensation for noisy speech emotion recognition

Proposed emotion recognition system for noisy speech is depicted in Figure 1, which consists of a feature extraction at the front end, followed by a conventional classifier. We propose the feature compensation and selection 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,

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

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 Equation 1 as,

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

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.

A. Feature compensation

In this paper, we propose to use the Vector Taylor Series (VTS) expansion and the VTS with psychoacoustic corruption function as the feature compensation, which are then used to compensate Gaussian mixture model (GMM) trained with the clean speech. 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], [13], 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 DCT matrix and it's inverse respectively. On the other hand, $\vec{y}$, $\vec{x}$, $\vec{h}$ and $\vec{n}$ are the Mel Frequency Cepstral Co-efficients (MFCC) domain distorted speech, clean speech, channel factor, and additive noise parameters. But, according to psychoacoustic corruption model [14], only the portion of noise which is above the masking threshold of clean speech is added to the speech. The psychoacoustic corruption function (as described in [15]), is used to modify the Equation 3 by incorporating the auditory masking criteria (i.e. VTS-AM), as the following,

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

where $\vec{w}$ is the scaling factor, which depends on the masking threshold of clean speech. Compensated model parameters can be computed by following similar methods as described in [13], [16]. The modified Taylor series component G which is the Jacobian of the mismatch function is defined as:

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

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:

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

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. $\vec{\mu}_y$ and Σ_y are the compensated mean and variance. In this approach, a GMM is trained on the clean speech denoted as $\lambda_x = \{\vec{\mu}_x, \vec{\sigma}_x, \vec{w}\}$. Next, the GMM parameters (mean and variance) are compensated according to the method described in [11]. Let the compensated model be denoted as $\lambda_y = \{\vec{\mu}_y, \vec{\sigma}_y, \vec{w}\}$. The pseudo-clean features $\vec{x}_{MMSE}$ are estimated from the noisy observations as [17]:

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

where $\vec{o}$ is the noisy speech features. $p(\vec{o}|\lambda_{ym})$ is the posterior probability for the m^{th} Gaussian mixture component of the

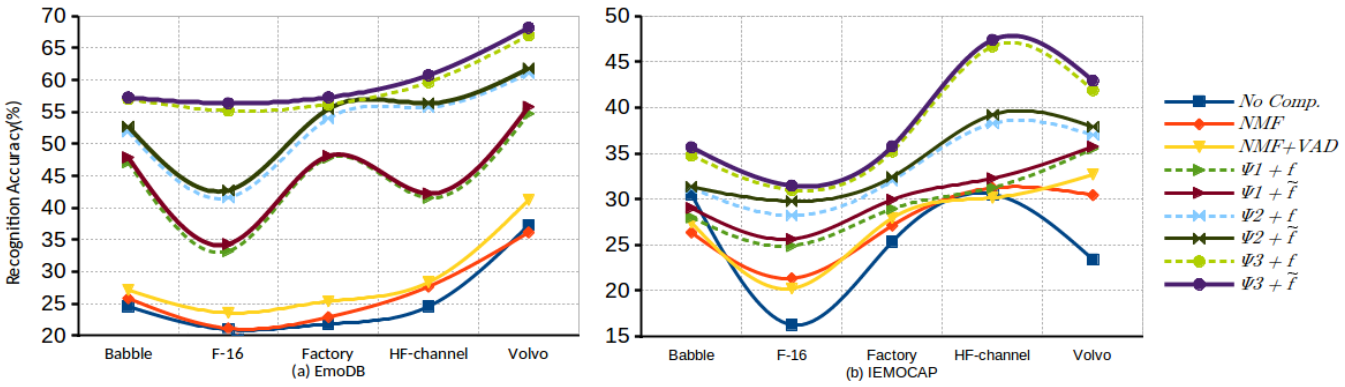

Fig. 2. Mean recognition accuracy (%) for different systems for (a) EmoDB (b) IEMOCAP. No Comp: no compensation, $\Psi 1$: log-MFCC + VTS, $\Psi 2$: log-MFCC + VTS-AM, $\Psi 3$: 10^{th} root-MFCC + VTS-AM, f: without feature selection, $\widetilde{f}$:(non-MFCC) feature selection

noise compensated GMM against the observation $\vec{o}$. $\vec{\mu}_{ym}$ is the m^{th} component of the noise compensated GMM and $\vec{\mu}_{xm}$ is the m^{th} component of the clean GMM.

1) Log and root compression: In MFCC feature computation, use of log-compression on the mel-filter bank energies has been the practice. The purpose of applying logarithm is to reduce the dynamic range of the feature and also to make data less sensitive towards variability [11], [18]. However, root compression is the other alternatives for reducing dynamic range of the mel-filter bank features. The benefit of logarithmic function is that channel effect can be discarded through cepstral mean and variance normalization, which is not possible for root compression. On the other hand, root compression exhibits superior noise robustness, which might yield better compaction of the spectral energy than other compression techniques.

2) High-level audio descriptors and feature selection: In the next stage of feature extraction process, we extract high-level descriptors (i.e. statistical functionals) from all low-level extracted features. Since the compensation is followed for MFCC features only, we have used a robust feature selection algorithm (for other low-level descriptors) that selects the relevant features in diverse noise conditions.

We used Information Gain Ratio based feature selection (IGR-FS) for feature reduction, where highly relevant attributes are found by their entropy, and ranking of attributes is independent of the classifier [2]. We also tried correlation based feature selection (CFS) [19]. 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 non-compensated (i.e. non-MFCC) features was found to give the best performance.

III. Experiments

A. Database and experimental set-up

We experimented with 2 standard emotional databases, namely (1) Berlin emotional database (EmoDB) [20], [21] and (2) Interactive emotional dyadic motion capture database (IEMOCAP) [22], [23], which were contaminated by noise to test our proposed techniques. Emo-DB 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 [24]. FANT toolkit is used for contamination of noise to clean speech at 5 SNR levels (0dB ,5dB, 10dB, 15dB, and 20dB) [25].

In all our experiments, we extracted 6 types of LLDs (frame-length of 20 ms and frame-shift of 10 ms), namely, *log*-energies, voice probability, frequency-band energies, F0, ZCR, and MFCC. Moreover, we took 39 statistical functionals (up to fourth order) for all the LLDs. We have used openSMILE toolkit for extracting acoustic features other than the MFCCs [26]. 23 dimensional MFCC features (with Δ and $\Delta\Delta$) were extracted for feature compensation using Kaldi speech recognition toolkit [27]. 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. We use 10^{th} root since it has been mentioned in [11] that 10^{th} root provides best performance compared to other roots. It should be noted that all methods compared are unsupervised methods and hence a separate development set is not required for parameter tuning.

B. Results and Analysis

All our experiments have been conducted on *clean*-training and *noisy*-testing, which is a mismatched scenario. For all our experimentations, we followed 5 cross-validation (CV) setup, by splitting up the dataset into 5 sets (80%-20% for train-test), and following leave one noisy set out for testing in each validation. Mean of recognition accuracies (RA in %) over all the SNR levels for each type of noises are shown in Figure 2 for both the databases. It is clear from the performance plot that we always get substantial improvements by just using either the VTS or the VTS-AM compensation techniques

TABLE I
CATEGORICAL EMOTION RA (IN %) (5 TYPES OF NOISE WITH 5 SNR LEVELS) FOR DIFFERENT SYSTEMS: NO COMP: NO COMPENSATION, $\Psi 1$: LOG-MFCC + VTS, $\Psi 2$: LOG-MFCC + VTS-AM, $\Psi 3$: 10^{th} ROOT-MFCC + VTS-AM, $\widetilde{f}$: (NON-MFCC) FEATURE SELECTION

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

(dotted lines in the plot) in comparison with the systems that use NMF or NMF+VAD. This shows the tremendous potential of these techniques in speech emotion recognition. It is also evident that VTS-AM performs much better than VTS and it follows the trends reported for speech recognition tasks [11]. 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.

Emotion recognition accuracies (RA in %) for the two datasets and for different techniques are tabulated in Table I. Due to space constraint we could not include all the techniques from Figure 2 in this table, but major results are covered. For Emo-DB, we observed absolute improvements (mean of RA (in %) across 5 SNR levels) of 2.58 (Babble), 2.64 (F-16), 3.52 (Factory), 3.78 (HF-channel) and 3.94 (Volvo) when NMF-VAD is used over the baseline (i.e. no compensation). Note that the performance of NMF-VAD was better than using only VAD or only NMF as observed in our previous work [3], [4], and similar trend is found here even for mismatched train-test scenario (also see Figure 2). While using both feature compensation and selection, overall trend found in terms of performances is: $\Psi 3$ (i.e. 10^{th} root-MFCC+VTS-AM) $>$ $\Psi 2$ (i.e. log-MFCC+VTS-AM) $>$ $\Psi 1$ (i.e. log-MFCC+VTS) (Table I and Figure 2). For both compensated and non-compensated MFCC features, we extract the HLDs. However, best results are found when feature selection (i.e. $\widetilde{f}$ using IGR-FS) is used for HLDs of non-MFCC features together with feature compensation. Best absolute improvements (mean of RA (in %) across 5 SNR levels) of 32.73 (Babble), 35.46 (F-16), 35.48 (Factory), 36.21 (HF-channel) and 30.91 (Volvo) are observed for $\Psi 3$ (10^{th} root-MFCC+VTS-AM)+$\widetilde{f}$) over the baseline. Similarly, absolute improvements (mean of RA (in %)) of 23.35 (Babble), 13.41 (F-16), 26.31 (Factory), 17.69 (HF-channel) and 18.53 (Volvo) are observed for $\Psi 1$ (log-MFCC+VTS)+$\widetilde{f}$) over the baseline. And, absolute improve-

ments (mean of RA (in %)) are observed as 28.17 (Babble), 21.82 (F-16), 33.64 (Factory), 31.82 (HF-channel) and 24.54 (Volvo) $\Psi 2$ (log-MFCC+VTS-AM)+$\widetilde{f}$) over the baseline.

Similarly for IEMOCAP database, absolute improvements (mean of RA (in %)) of 3.97 (F-16), 2.59 (Factory), and 9.32 (Volvo) were observed when NMF-VAD is used over the baseline. However, no improvements (RA (in %)) are observed for (5dB, 10dB, 15dB) of Babble and (5dB) of HF-channel noisy speech, using NMF-VAD w.r.t. no compensation baseline. However, best improvements (mean of RA (in %)) of 5.15 (Babble), 15.23 (F-16), 10.44 (Factory), 16.9 (HF-channel) and 19.57 (Volvo) are observed for $\Psi 3$ (10^{th} root-MFCC+VTS-AM)+$\widetilde{f}$) over the baseline (as shown in Table I and Figure 2). And, the next best system with absolute improvements (mean of RA (in %)) are observed as 0.9 (Babble), 13.53 (F-16), 7.09 (Factory), 8.7 (HF-channel) and 14.52 (Volvo) $\Psi 2$ (log-MFCC+VTS-AM)+$\widetilde{f}$) over the baseline. While experimenting with system ($\Psi 1$ (log-MFCC+VTS)+$\widetilde{f}$), absolute improvements (mean of RA (in %)) of 9.38 (F-16), 4.6 (Factory), 1.75 (HF-channel) and 12.37 (Volvo) are observed over the on compensation baseline, but no improvements are found for 10dB and 15dB of Babble noise contaminated speech. However, it should be noted that the performance of ($\Psi 1$ (log-MFCC+VTS)+$\widetilde{f}$) is still better than the NMF or NMF+VAD system. From these experiments, it is clear that the best performance is achieved by applying VTS-AM feature compensation along with 10^{th} root and feature selection for non-MFCC high level descriptors.

IV. Conclusion

In this paper, we propose front-end feature compensation and selection technique for noisy speech emotion recognition. VTS based feature compensation with psychoacoustic masking has been proved to be beneficial. While computing MFCC features, 10^{th} root compression found to gel well with VTS-AM in comparison with the log compression. Concatenation of selected high-level descriptors of low-level non-MFCC features and VTS-AM compensated MFCCs yielded the best performance in train-test mismatched conditions. The fact that the proposed method outperforms previously used NMF-based enhancement or even the NMF-VAD by a significant margin, clearly indicates its efficacy.

References

[1] M. You, C. Chen, J. Bu, J. Liu, and J. Tao, "Emotion recognition from noisy speech," *IEEE ICME*, 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] Meghna Pandharipande, Rupayan Chakraborty, Ashish Panda, and Sunil Kumar Kopparapu, "An unsupervised frame selection technique for robust emotion recognition in noisy speech," *Eusipco*, 2018.

[4] Meghna Pandharipande, Rupayan Chakraborty, Ashish Panda, and Sunil Kumar Kopparapu, "Robust front-end processing for emotion recognition in noisy speech," *ISCSLP*, 2018.

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

[6] Mingyu You, Chun Chen, Jiajun Bu, Jia Liu, and Jianhua Tao, "Emotion recognition from speech signal combining pca and lda," *Emotion Recognition, ICME*, 2006.

[7] Lijiang Chen, Xia Mao, Yuli Xue, and Lee Lung Cheng, "Speech emotion recognition: Features and classification models," *Digital Signal Processing*, 2012.

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

[9] C. Huang, G. Chen, Hua Yu, Y. Bao, and Li Zhao, "Speech emotion recognition under white noise," *Archives of Acoustics*, 2013.

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

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

[12] Felix Weninger, Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi, "Recognition of nonprototypical emotions in reverberated and noisy speech by nonnegative matrix factorization," *EURASIP Journal on Advances in Signal Processing*, 2011.

[13] Jinyu Li, Li Deng, Dong Yu, Yifan 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.

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

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

[16] Acero Alex, Deng Li, Kristjansson Trausti, and Zhang Jerry, "HMM adaptation using vector taylor series for noisy speech recognition," *Proc. Int. Conf. on Spoken Language Processing*, 2000.

[17] 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.

[18] Sourabh Ravindran, David V. Anderson, and Malcolm Slaney, "Improving the noise-robustness of mel-frequency cepstral coefficients for speech processing," *SAPA@INTERSPEECH*, 2006.

[19] M. A. Hall, *Correlation-based Feature Subset Selection for Machine Learning*, Ph.D. thesis, University of Waikato, Hamilton, New Zealand, 1998.

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

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

[22] C. Busso, M. Bulut, Chi-Chun Lee, Abe Kazemzadeh, E. Mower, S. Kim, Jeannette N. Chang, S. Lee, and S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, 2008.

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

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

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

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

[27] D. Povey, A. Ghoshal, G. Boulianne, O. Glembek 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.