ROBUST FRONT-END PROCESSING FOR SPEECH RECOGNITION IN NOISY CONDITIONS

Biswajit Das, Ashish Panda

TCS Innovation Labs, Mumbai
Yantra Park, Thane, Maharashtra, India, 400601.
Email: b.das@tcs.com, ashish.panda@tcs.com

ABSTRACT

In this paper, we investigate the applicability and effectiveness of advanced feature compensation techniques in devising a robust front-end for Automatic Speech Recognition (ASR). First, the Vector Taylor Series (VTS) equations are altered by bringing in the auditory masking factor. The resultant VTS approximation is used to compensate the parameters of a clean speech model and a Minimum Mean Square Error (MMSE) estimate is used to estimate the clean speech features from noisy features. Second, we apply root-compression instead of conventional log-compression to the mel-filter banks energy. Third, we apply a frame selection method to eliminate the noise dominated frames to improve the performance in high noise scenarios. The proposed algorithms are validated on noise corrupted Librispeech and TIMIT speech recognition databases and are shown to provide significant gain in performance.

Index Terms— Noise robust speech recognition, Auditory Masking, Vector Taylor series, Root Compression, Frame Suitability Measure.

1. INTRODUCTION

The performance of the ASR systems degrade significantly due to adverse conditions in the test environment. Hence, incorporation of ASR system in real life applications remains limited. To mitigate the effect of noise in ASR systems, different approaches have been proposed. In feature domain, Cepstral Mean and Variance Normalization technique is applied on the top of Mel-Frequency Cepstral Coefficient (MFCC) features to deal with channel mismatch. Besides this, other kind of features like, auditory based modulation spectral feature for reverberant noise [1] and deep belief network based tandem features [2] have been employed for noise robust ASR. Weighted denoising auto-encoder based on Weiner filter has also been investigated for noise robust speech recognition [3]. Auditory feature based on gammatone filter [4] has been explored as an alternative to the MFCCs. Other than this, signal pre-processing technique like, non-negative matrix factorization method [5, 6] has been attempted for removing noise from the signal.

Deep recurrent denoising autoencoder (DRADE) is a process of extracting clean features from the noisy features [7]. The DRADE technique outperformed SPLICE based denoising method in [8]. But, the DRADE demands lot of stereo data (clean and noisy). Different types of noise robust features like normalized modulation coefficients (NMC), modulation of medium duration speech amplitudes (MMeDuSA) and Damped Oscillator Coefficients (DOC) have been attempted in [9] for noise robust speech recognition. Different model domain techniques have been proposed for Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) architecture like Parallel Model Combination (PMC) [10], Vector Taylor Series (VTS) expansion [11] and Psychoacoustic Model Compensation (Psy-Comp) [12, 13, 14]. But, these techniques are incompatible with the Deep Neural Network (DNN) based speech recognition systems. So, one alternative is to apply these techniques in the front-end processing as suggested in [15, 16].

In [17], we proposed a Vector Taylor Series with Auditory Masking (VTS-AM), which consistently outperformed the conventional VTS method. In this paper, we propose further improvements in the front-end processing. We investigated the effectiveness of the root-compression [18] instead of the conventional log-compression of the mel-filter banks energy. We also propose a method of using the root-compression in conjunction with the VTS and the VTS-AM feature compensation. To the best of our knowledge, this has not been attempted yet in the literature. We show that the root compression improves the performance of both these methods by a significant margin. To further improve the performance in low Signal-to-Noise-Ratio (SNR) scenarios, we also employ a spectral variance based frame selection method.

The remainder of the paper is organized as follows. Section 2 briefly describes the compression techniques for front-end processing and frame suitability measure (FSM) is discussed in Section 3. Section 4 describes the Vector Taylor Series expansion with Auditory Masking. Section 5 describes the estimation of enhanced features, while the overall algo-

rithm is presented in Section 6. Section 7 and 8 deal with the experiments and results and Section 9 concludes this paper.

2. 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 [19]. We can also use root compression instead of logarithm compression for reducing dynamic range of the mel-filter bank features. Relation between root function and logarithm function is reported in [18] as follows:

$$f_r(x) = \frac{x^r - 1}{r} \tag{1}$$

If we expand the Equation 1 using Taylor series, it will be related to logarithm function as follows:

$$\lim_{r \to 0} f_r(x) = \log x \tag{2}$$

The befefit of logarithm function is that channel effect can be discarded through cepstral mean and variance normalization, which is not possible for root compression. Nonetheless, empirical evidence shows that root compression exhibits noise robustness for ASR applications [19, 4]. The reason behind improved performance, as has been justified in [19], is that root compression may result in better compaction of the spectral energy.

3. FRAME SUITABILITY MEASURE

Frame suitability measure (FSM) approach helps to select appropriate frames for improved ASR performance. Energy Normalized Variance has been investigated in [20] for identifying noisy speech frames. Energy-normalized variance can be defined as follows:

$$N\_VAR = \frac{\sum_{i=0}^{N}(X_i - \bar{X})^2}{\sum_{i=0}^{N}(X_i)^2} \tag{3}$$

where $\bar{X}$ is the mean of mel-filter banks energy and N is the total number of mel-filter banks. Values of N_VAR are in the range of 0 to 1. High values of N_VAR represent speech regions while lower values represent noise dominated regions. After computing N_VAR for all frames, we can select the frames with high values of N_VAR.

4. TAYLOR SERIES EXPANSION WITH AUDITORY MASKING

Traditional assumption of noise corruption model is that the speech and noise are additive in the spectral magnitude domain. But, according to psychoacoustic corruption model [13], human beings perceive only the portion of noise which is above the masking threshold of clean speech and only the perceived noise is added to the speech. The psychoacoustic corruption function is, as described in [12, 14], defined as:

$$Y_f = X_f + N_f - 10^{\frac{T_{mf}}{20}} \tag{4}$$

where Y_f is the corrupted speech signal in the mel-filter bank domain. X_f and N_f are the clean speech and the additive noise respectively. f denotes the mel-filter index. T_{mf} is the masking threshold of the clean speech X_f. Masking thresholds of clean speech is computed as follows [14]:

$$T_{mf} = 20\log_{10}(X_f) - 0.275.C_f - 6.025 \quad \text{(dB)} \tag{5}$$

where C_f is the central frequency of the mel-filter in bark scale.

Considering the channel factor H in the corruption function 4, it can be written as:

$$Y_f = H_f X_f + N_f - 10^{\frac{T_{mf}}{20}} \tag{6}$$

After re-arranging the above Equation 6, we can redefine the corruption as follows:

$$Y_f = W_f X_f H_f + N_f \tag{7}$$

where H_f is the channel factor and N_f is the additive noise. Here, W_f is the weighting factor which can be expressed as follows:

$$W_f = \frac{X_f - 10^{\frac{T_{mf}}{20}}}{X_f} \tag{8}$$

Now, if we re-arrange the Equation 7 after taking log and multiplying with the discrete cosine transform (DCT) matrix, we can get a non-linear distortion model in the cepstral domain as:

$$\vec{y}^s = \vec{x}^s + \vec{h}^s + \vec{w}^s + Clog(1 + exp(C^{-1}(\vec{n}^s - \vec{x}^s - \vec{h}^s - \vec{w}^s))) \tag{9}$$

where C and C^{-1} is the DCT matrix and it's inverse respectively. On the other hand, $\vec{y}$, $\vec{x}$, $\vec{h}$, $\vec{w}$, and $\vec{n}$ are distorted speech, clean speech, channel factor, scaling factor and additive noise respectively and all these parameters are in MFCC domain. We can compute the compensated model parameters following similar methods described in [11, 21].

The modified Taylor series component G which is the Jacobian of the mismatch function is defined as:

$$G = C \bullet diag\left(\frac{1}{1 + exp(C^{-1}(\vec{\mu_n} - \vec{\mu_x} - \vec{w} - \vec{h}))}\right) \bullet C^{-1} \tag{10}$$

It is important to note that component G is derived using only the static portion of model mean and noise mean. In the next step, we can compensate the model mean and variance as follows:

$$\vec{\mu}_y = \vec{\mu}_x + \vec{h} + \vec{w} + Clog(1 + exp(C^{-1}(\vec{\mu_n} - \vec{\mu}_x - \vec{w} - \vec{h}))) \tag{11}$$

and

$$\Sigma_y \approx G\Sigma_x G^T + (I - G)\Sigma_n (I - G)^T \quad (12)$$

where I and T are the identity matrix and transpose respectively. $\vec{\mu}_y$ and Σ_y are the compensated mean and variance. Equations 10 and 11 are different from the traditional VTS expansion methods, discussed in [11, 21]. Here, we have introduced the weighting factor $\vec{w}$, which depends on the masking threshold of clean speech.

5. ENHANCED FEATURE ESTIMATION

In the Section 4, we described the clean GMM model parameters compensation method using estimated noise statistics. Here, we estimate the pseudo-clean features from noisy features by the MMSE method as described in [16]. In this approach, we need a GMM, which is trained on clean speech and let it be denoted as $\lambda_x = \{\vec{\mu}_x, \vec{\sigma}_x, \vec{w}\}$. Next, the GMM parameters (mean and variance) are compensated according to Section 4. Let the compensated model be denoted as $\lambda_y = \{\vec{\mu}_y, \vec{\sigma}_y, \vec{w}\}$. The pseudo-clean features are estimated from the noisy observations by first order VTS approximation as

$$\begin{aligned} \vec{x}_{MMSE} &= E(\vec{x}|\vec{o}) = \int \vec{x} p(\vec{x}|\vec{o}) dx \\ &= \vec{o} - \sum_{m=0}^{M-1} p(\vec{o}|\lambda_{ym})(\vec{\mu}_y - \vec{\mu}_{xm}) \end{aligned} \quad (13)$$

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 noise compensated GMM against the observation $\vec{o}$. $\vec{o}$ is the noisy speech signal. The $\vec{\mu}_{ym}$ is the $\vec{m}^{th}$ component of the noise compensated model. On the other hand, $\vec{\mu}_{xm}$ is the $\vec{m}^{th}$ component of the clean model.

6. PROPOSED METHOD

We will describe the proposed methods step by step. For this proposed method, we need a GMM which is trained using clean training feature.

1. Estimate initial noise statistics (noise mean and variance), $\vec{\mu_n}$ and Σ_n using starting and ending frames.
2. Compensate GMM means and variances using Equations 11 and 12.
3. Re-estimate the additive noise $\vec{\mu_n}$ and channel factor $\vec{\mu_h}$ using Expectation Maximization (EM) algorithm described in [11].
4. Compensate clean GMM means and variances using Equations 11 and 12 with re-estimated noise statistics.
5. Now, the pseudo-clean features are estimated from noisy feature using Equation 13 in MFCC domain.
6. Next, we convert enhanced MFCC to mel-filter banks energy and apply the frame selection method using Equation 3 to eliminate noise dominated frames .
7. Then, we apply the 10^{th} root or log-compression, according to the experimental need.
8. The compressed mel-filter banks energy are used for recognition in DNN architecture.

7. EXPERIMENTAL SETUP

We have conducted all the experiments on two separate speech recognition corpora: TIMIT and Librispeech [22]. Kaldi Speech Recognition Toolkit [23] has been used for all experiments. For acoustic model training, we used only clean speech waveforms. To prepare test data of both databases, we corrupted clean test waveforms with different noise types like hfchannel (HF), F-16 and babble at various SNRs like 0dB, 5dB, 10dB and 15dB. To accomplish this task, we have used the standard Filtering and Noise Adding Tool (FaNT) [24].

For both databases, we extracted mel-filter bank features with logarithm and 10^{th} root compression. We built two separate acoustic models with training data of two databases. In case of TIMIT database, we followed standard training recipe of Kaldi toolkit. In this recipe, CMN technique is used for feature normalization. Moreover, only two hidden layers are considered for DNN structure. In the other experiment for Librispeech database, we selected approximately 30 hours of data from the database. It consists of total 585 speakers out of which 284 male and 301 female speakers. In this setup, we adopted the Librispeech recipe of Kaldi toolkit. As it is large database, we selected four hidden layers for DNN architecture. In this experiment, we also built two separate language model from the training text of the databases.

To train the clean speech GMM (128 components), we considered 23 dimensional static MFCC features. It helped transform the model parameters to mel-filter bank domain without any approximation, since we are using 23 mel-filters to compute the MFCC features. Noisy features are compensated using VTS and the VTS-AM. After that, we transformed the compensated feature from MFCC domain to mel-filter bank domain. Next, we applied logarithm and 10^{th} root compression on the VTS compensated features and those are called "VTS_log" and "VTS_root" features respectively. Similarly, we also used both logarithm and root compression with the VTS-AM compensated features and we denoted the enhanced features as "VTS-AM_log" and "VTS-AM_root", respectively.

In the final step, we applied frame selection technique on the VTS-AM_root. It should discard the noise dominated

frames to improve the ASR performance. While frame selection technique has not been used with other methods, we propose to do so in the near future.

8. RESULTS

In the baseline experiment of TIMIT database, we have not applied any feature enhancement technique. The baseline system with log compressed features are denoted as "BASE_log" and with the root compressed features are denoted as "BASE_root". Phoneme Error Rate (PER) for "BASE_log" and "BASE_root" are 22.7% and 22.8% respectively with clean test data. Table 1 shows experimental results for different noise robustness techniques for TIMIT database. It can be observed that phoneme recognition performance drastically degrades with the presence of different types of noises. Experimental results shows 6% absolute performance gain using only root-compression over the log-compression. VTS as well as VTS-AM techniques perform very well along with log-compression, with VTS-AM consistently outperforming the VTS method. Moreover, significant performance gain achieved after combining root-compression with VTS and VTS-AM techniques and the performance gap between VTS and VTS-AM narrows considerably. With root-compression, there is very little to choose between VTS and VTS-AM. The frame selection method helps in improving performance for the lower SNR scenarios only, while the the higher SNR scenarios do not benefit from it. This may be because in higher SNR scenarios, there are fewer noise dominated frames.

		Baseline		VTS		VTS-AM		VTS-AM and FSM
	SNR	log	root	log	root	log	root	root
HF	0dB	79.6	70.0	60.5	58	59.3	58.4	57.6
	5dB	64.2	56.0	56.0	48.4	49.8	48.7	48.2
	10dB	47.6	42.1	41.9	40.2	40.8	40.1	40
	15dB	36.3	33.7	37.3	32.9	33.1	33	33.1
F-16	0dB	89.5	77.3	67.4	62.4	64.7	62.6	62.1
	5dB	78.2	65.3	56	52.1	54.3	52.2	51.8
	10dB	58.5	49.9	46.9	41.9	44.2	41.9	42
	15dB	42.1	38.2	37.3	35	35.3	34.9	34.8
BABBLE	0dB	82.3	75.6	71.6	67	68.7	67.6	65.9
	5dB	68.3	64.0	57.4	53.5	55.8	54.4	54
	10dB	52.5	50.3	46.2	43.4	44.4	43.4	43.3
	15dB	40.4	39.2	38.0	35.6	36.6	35.6	35.4
Average		61.6	55.1	51.3	47.5	48.9	47.7	**47.3**

Table 1. Phoneme Error Rate (PER) of different techniques with different SNR level for TIMIT

Table 2 shows experimental results of different noise robust techniques on Librispeech database. we have got 13.82% and 14.07% Word Error Rate (WER) for "BASE_log" and "BASE_root" respectively with clean test speech. We can observe that the performance of the various methods follow the same trend as in the TIMIT experiments. It is interesting to observe that auditory masking in Taylor series exhibits significant improvement over traditional VTS technique for log-compression. However, the with root-compression, both VTS and VTS-AM perform almost at the same level. This opens up possibilities that a more suitable formulation of auditory masking might be beneficial for the root compressed features. The frame selection method used in this work also indicates that better frame selection methods might aid the performance of the feature enhancement techniques.

		Baseline		VTS		VTS-AM		VTS-AM and FSM
	SNR	log	root	log	root	log	root	root
HF	0dB	83.92	77.04	72.58	68.62	70.13	68.46	68.16
	5dB	62.25	52.2	49.14	45.16	47.55	44.89	45.18
	10dB	36.68	31.09	31.85	29.31	31.33	29.34	29.56
	15dB	22.95	21.87	22.27	21.59	22.03	21.57	21.82
F-16	0dB	90.42	86.75	80.75	74.67	76.57	74.77	74.05
	5dB	72.8	61.53	56.92	48.92	51.57	48.99	48.87
	10dB	43.72	35.38	34.39	29.91	32.05	29.73	29.76
	15dB	24.31	22.48	22.29	20.73	21.38	20.72	20.76
BABBLE	0dB	85.94	84.18	83.49	80.07	81.15	80.51	79.58
	5dB	61.85	58.36	59	53.69	56.53	54.19	53.73
	10dB	34.79	32.63	34.25	31.58	33.18	31.57	31.63
	15dB	22.14	20.84	21.55	20.77	21.05	20.85	20.68
Average		53.48	48.70	47.37	43.55	45.17	43.8	**43.64**

Table 2. Word Error Rate (WER) of different techniques with different SNR level for Librispeech

9. CONCLUSION

In this paper, we have proposed robust front end methods for speech recognition in noisy conditions. We have shown that using root compression in conjunction with VTS or VTS-AM methods of feature enhancement can greatly improve the performance. Also, employing a simple frame selection method based on energy normalized variance can aid the performance in low SNR scenarios. Currently, we are looking into better frame selection methods and also a formulation of auditory masking that might be more suitable for root compressed features.

10. REFERENCES

[1] H. K. Maganti and M. Matassoni, "An auditory based modulation spectral feature for reverberant speech recognition," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[2] O. Vinyals and S. V. Ravuri, "Comparing multilayer perceptron to deep belief network tandem features for robust asr," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 4596–4599.

[3] B. Xia and C. Bao, "Wiener filtering based speech enhancement with weighted denoising auto-encoder and

noise classification," in *Speech Communication*, 2014, vol. 60, pp. 13 – 29.

[4] R. Schluter, I. Bezrukov, H. Wagner, and H. Ney, "Gammatone features and feature combination for large vocabulary speech recognition," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, April 2007, vol. 4, pp. IV–649–IV–652.

[5] K. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," 2008.

[6] F. Weninger, M. Wllmer, J. Geiger, B. Schuller, J. F. Gemmeke, A. Hurmalainen, T. Virtanen, and G. Rigoll, "Non-negative matrix factorization for highly noise-robust asr: To enhance or to recognize?," pp. 4681–4684, March 2012.

[7] A.L. Maas, Q.V. Le, T.M. ONeil, O. Vinyals, and P. Nguyen, "Recurrent neural networks for noise reduction in robust asr," in *in Proc. Interspeech. ISCA*, 2012.

[8] L. Deng, A. Acero, L. Jiang, J. Droppo, and X. Huang, "High-performance robust speech recognition using stereo training data," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 301–304, 2001.

[9] V. Mitra, W. Wang, H. Franco, Y. Lei, C. Bartels, and M. Graciarena, "Evaluating robust features on deep neural networks for speech recognition in noisy and channel mismatched conditions.," in *INTERSPEECH*, 2014, pp. 895–899.

[10] L. P. Wong and M. Russell, "Text-dependent speaker verification under noisy conditions using parallel model combination," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 457–460, 2001.

[11] 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," *Automatic Speech Recognition Understanding, 2007. ASRU. IEEE Workshop on*, pp. 65–70, Dec 2007.

[12] B. Das and A. Panda, "Psychoacoustic model compensation for robust continuous speech recognition in additive noise," *2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pp. 511–515, Dec 2015.

[13] A. Panda and T. Srikanthan, "Psychoacoustic model compensation for robust speaker verification in environmental noise," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 3, pp. 945–953, 2012.

[14] A. Panda, "A fast approach to psychoacoustic model compensation for robust speaker recognition in additive noise," *in Proc. INTERSPEECH. ISCA*, pp. 205–209, 2015.

[15] P. J. Moreno, B. Raj, and R. M. Stern, "A vector taylor series approach for environment-independent speech recognition," *IEEE International Conference on Acoustics,Speech and Signal Processing (ICASSP)*, vol. 2, pp. 733–736 vol. 2, May 1996.

[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," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7408–7412, May 2013.

[17] B. Das and A. Panda, "Vector taylor series expansion with auditory masking for noise robust speech recognition," in *The 10th IEEE International Symposium on Chinese Spoken Language Processing*, 2016.

[18] F. Hilger and H. Ney, "Evaluation of quantile based histogram equalization in combination with different root functions," in *FORTSCHRITTE DER AKUSTIK*, 2005, vol. 31, p. 225.

[19] S. Ravindran, D. V. Anderson, and M. Slaney, "Improving the noise-robustness of mel-frequency cepstral coefficients for speech processing," *Reconstruction*, vol. 12, pp. 14, 2006.

[20] K. Manohar and P. Rao, "Speech enhancement in nonstationary noise environments using noise properties," *Speech Communication*, vol. 48, no. 1, pp. 96–109, 2006.

[21] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "Hmm adaptation using vector taylor series for noisy speech recognition," in *International Conference on Spoken Language Processing (ICSLP)*, October 2000.

[22] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," pp. 5206–5210, April 2015.

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

[24] H. G. Hirsch and H. Finster, "The simulation of realistic acoustic input scenarios for speech recognition systems," in *International Conference on Spoken Language Processing (ICSLP)*, 2005.