INTEGRATING DENOISING AUTOENCODER AND VECTOR TAYLOR SERIES WITH AUDITORY MASKING FOR SPEECH RECOGNITION IN NOISY CONDITIONS

A. Biswajit Das, *

INRIA Bordeaux Sud-Ouest (GeoStat team)
Talence, France
Email: biswajit.das@inria.fr

Ashish Panda

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

ABSTRACT

We propose a new front-end feature compensation technique to improve the performance of Automatic Speech Recognition (ASR) systems in noisy environments. First, a Time Delay Neural Network (TDNN) based Denoising Autoencoder (DAE) is considered to compensate the noisy features. The DAE provides good gain in performance when it has been trained using the noise present in the test utterances ("seen" conditions). However, if the noise present in the test utterance is different to what was used in the training of the DAE ("unseen" conditions), then the performance degrades to a great extent. To improve the ASR performance in such unseen conditions, a model compensation technique, namely the Vector Taylor Series with Auditory Masking (VTS-AM) is used. We propose a new Signal-to-Noise Ratio (SNR) based measure, which can reliably choose the type of compensation to be used for best performance gain. We show that the proposed technique improves the ASR performance significantly on noise corrupted TIMIT and Librispeech databases.

Index Terms— Noise robust speech recognition, Auditory masking, Vector Taylor series, Time delay neural network, Denoising autoencoder.

1. INTRODUCTION

Inspite of all the advances in the ASR performance in the recent years, noisy speech is still a challenge. Different approaches have been reported in literature to improve the noise robustness of ASR systems. Feature normalization, such as cepstral mean and variance normalization, is widely used to deal with the speech degradation. Different feature extraction processes 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. A psychophysically inspired amplitude modulation filter bank based feature extraction scheme has been proposed in [3]. Different compression techniques [4, 5], such as root compression instead of log compression of the mel-filter energy, have improved ASR performance.

Several techniques like Vector Taylor Series (VTS) [6] and Psychoacoustic Model Compensation (Psy-Comp) [7, 8] have been proposed for Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) based techniques. These techniques have also been used successfully in the front-end processing for Deep Neural Network (DNN) based techniques in [9, 10]. The VTS-AM feature enhancement [4] has been shown to outperform the traditional VTS technique. Beside this, a deep Convolutional Neural Networks (CNNs) based noise robust speech recognition is proposed in [11], which outperforms long short-term memory Recurrent Neural Networks (RNNs) [12]. For robust feature computation, DAE [13] has provided significant improvement over other techniques. DAEs based on different types of networks like DNN, CNN, RNN [14, 15], and TDNN [16] have been investigated to improve robustness against noise.

In this paper, we propose a robust front-end for speech recognition. The advantage of a robust front-end is that it does not depend on the specific architecture of the speech recognition engine. Therefore, it can be used for robustness across different architectures. To build a robust front-end, we take a closer look at the performance of the TDNN based DAE. The TDNN-DAE architecture is the same as proposed in [16]. In particular, we examine the performance of the DAE in "seen" and "unseen" conditions. The "seen" condition is where the noise encountered in the test utterance was used in the training of the DAE and the "unseen" condition is where the noise encountered in the test utterance was not used in the training of the DAE. This study is important, since it is extremely difficult, if not downright impossible, to train the DAE for all types of noise. We show that although the DAE performs very well in the seen conditions, it under-performs compared to the VTS-AM for the unseen conditions. Therefore, it would be beneficial to employ the VTS-AM method where the DAE fails. However, the challenge is to automatically identify utterances where the DAE has failed. To this end, we propose a new SNR based measure, which can reliably indicate the failure or success of the DAE.

*The first author performed the work while at TCS Innovation Labs-Mumbai.

ISBN 978-90-827970-1-5

The remainder of the paper is organized as follows. Section 2 briefly describes the TDNN based DAE for robust feature processing. Section 3 describes the Vector Taylor Series expansion with Auditory Masking. Section 4 describes the computation of the SNR based measure and the overall algorithm is presented in Section 5. Section 6 and 7 deal with the experiments and results, while Section 8 concludes this paper.

2. TIME DELAY NEURAL NETWORK BASED DENOISING AUTOENCODER

In TDNN architecture [17], network is organized with narrow contexts in initial layers and wider context for deeper layer to learn the transform. TDNN architecture is motivated and employed for DAE in [16] to estimate enhanced features. In this architecture, input features consist of noisy speech Mel-Frequency Cepstral Coefficients (MFCCs), whereas target features are the corresponding clean speech MFCCs. Back propagation training approach computes network parameters such that it can capture the feature enhancement mapping. We have followed the TDNN network architecture as described in the study [16]. This DAE network has 4 hidden layers and each hidden layer consists of 1024 ReLU activation nodes.

3. 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 [18], only the portion of noise which is above the masking threshold of clean speech is added to the speech. The psychoacoustic corruption function is described in [7, 8]. In [4], we have altered VTS equations by bringing in the auditory masking criteria, which is known as VTS-AM. 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 [4]. 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 [9]:

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

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

4. A NEW APPROACH FOR SNR COMPUTATION

The aim is to reliably identify conditions, where the DAE has failed to enhance the features. If the DAE has failed to enhance the features, then it can be expected that the SNR of the DAE enhanced features will be lower in unseen cases as compared to the seen cases. The SNR of the DAE enhanced features, then, can be a reliable indicator as to whether the DAE has failed to work. The challenge, however, is to compute the SNR of the DAE enhanced features, since we do not have access to either the clean features or the noise energy.

Leveraging on the noise-estimation algorithm of the VTS-AM method, we describe, here, a new method of computing the SNR of a signal without needing the clean signal energy or the noise energy. The VTS-AM method employs a GMM trained on the clean speech. Let it be denoted as $\lambda_x = \{\vec{\mu}_x, \vec{\sigma}_x, \vec{w}\}$. The GMM means $\vec{\mu}_x$ can be converted to mel-spectral domain through multiplication with Inverse Discrete Cosine Transformation (IDCT) matrix and exponentiation operation. Let the mel-spectral representation of the GMM means be denoted as $\vec{\mu}^e$. The mean $\vec{\mu}^{avg}$ of the GMM means can then be computed as:

$$\vec{\mu}^{avg} = \frac{\sum_i^M \vec{\mu}_i^e}{M} \qquad (2)$$

where M is the total number of mixture in GMM. The average energy of clean speech means can then be computed as: $S = \sum_d^D (\mu_d^{avg})^2$, where D is the dimension of the feature vector. The reason behind taking the average of the GMM means in the above equation is that it would provide an estimate of the average energy in a frame in the training set.

The VTS-AM technique estimates the noise energy using Expectation Maximization (EM) technique. Let us denote the noise vector estimated by VTS-AM for a given noisy utterance as $\vec{\mu}_n$ in MFCC domain. Next step, convert $\vec{\mu}_n$ to mel-filter bank domain which is denoted as $\vec{\mu}_n^e$. The noise energy can then be computed as: $N = \sum_d^D (\mu_{nd}^e)^2$.

Once we have the clean speech energy S and the noise energy N, we can compute the SNR as:

$$SNR = 10log_{10}\frac{S}{N} \qquad (3)$$

It is worth noting that the above described SNR is not the true SNR of the noisy signal. It would be more accurate to say that the SNR computed this way is a ratio of the average energy of the training speech to the noise in the noisy signal. However, since the average energy of the training speech is a constant, this SNR would indicate the relative noise content of a signal. In other words, if the SNR of signal A is higher than the SNR of the signal B, then it can be reliably said that signal A has lower level of noise content compared to signal B.

ISBN 978-90-827970-1-5 © EURASIP 2018

5. PROPOSED FRONT-END PROCESSING

In this section we outline the proposed method. First, a TDNN based DAE model is trained using noisy speech utterances and their corresponding clean speech utterances. In Figure 1, method of TDNN DAE training approach is described.

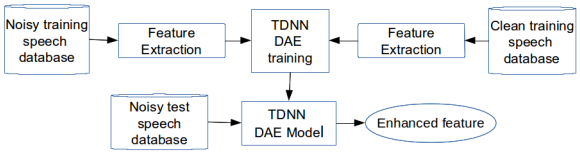

Fig. 1. TDNN DAE training

During the recognition phase, the test utterance is input to the DAE and enhanced (pseudo clean) features are computed. After feature enhancement, SNR is computed using proposed method as discussed in Section 4. If the SNR value is greater than the threshold, DAE enhanced features are used for final recognition. Otherwise, the test utterance is enhanced with VTS-AM technique discussed in Section 3. This scheme is illustrated in Figure 2.

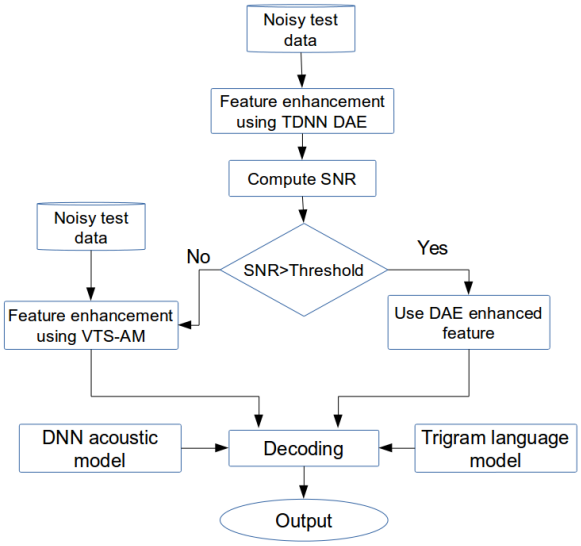

Fig. 2. Diagram of the proposed method

6. EXPERIMENTAL SETUP

To validate the effectiveness of the proposed method, we have considered two different speech databases TIMIT and Librispeech. All the experimental setups are implemented with Kaldi speech recognition toolkit [19]. To prepare the test data for both databases, we corrupted clean test waveforms with different noise types like hfchannel (HF), F-16 and babble (BAB) at various SNRs like 0dB, 5dB, 10dB and 15dB. To accomplish this task, we have used the Filtering and Noise Adding Tool (FaNT) [20]. More details on data preparation can be found in [4]. We have followed the DNN framework for speech recognition and we trained two sets of acoustic models for TIMIT and Librespeech databases using the clean training speech. 23 dimensional Mel Frequency Cepstral Coefficients are used as feature vectors.

Contexts for the DAE network with four hidden layers is organized as (-2,-1,0,1,2) (-1,2) (-3,3) (-7,2) (0) which is asymmetric in nature. Input temporal context for the network is set to (-13,9).

The TDNN DAE was trained on the training data of the respective databases, which was corrupted with the desired noise types. We used several different combination of noise types to simulate the seen and unseen conditions. For example, if the goal was to simulate a seen condition for the noise type HF, then the DAE was trained on training data corrupted with different levels of HF, F-16 and BAB noise. On the other hand, if the goal was to simulate an unseen condition for the noise type HF, then the training data was corrupted with different levels of F-16 and BAB noise, leaving out the HF noise. For VTS-AM technique, two seperate GMMs with 128 components are trained with clean training data from TIMIT and Librispeech database.

7. RESULTS

We have conducted a series of experiments to observe performance of various robustness techniques. Experimental result for the TIMIT database, in the terms of Phoneme Error Rate (PER) is provided in Table 1. The PER achieved for the clean test data was 22.7%. It can be observed that recognition accuracy degrades drastically according to the noise level. After employing VTS_AM technique 12% absolute improvement is observed on an average. Interesting results are observed for the DAE based feature enhancement. While 21% absolute improvement is obtained for seen conditions, for unseen conditions, the improvement is only about 3.4%. It can be clearly seen that VTS-AM can offer significantly better performance gain for unseen conditions.

Table 2 shows experimental results from Librispeech database. We have achieved 14.04% WER for clean test data of Librispeech. Here also, we can observe absolute 7.42% performance improvement for VTS-AM method and 20.4% absolute improvement for seen condition using TDNN based DAE technique. However, for unseen conditions, there is a performance degradation of about 6.8% compared to the system with no enhancement. This seems to indicate that the DAE based enhancement may not be suitable for all conditions.

ISBN 978-90-827970-1-5 © EURASIP 2018

	SNR	No Enhancement	VTS-AM	DAE seen condition	DAE unseen condition
F-16	0dB	87.6	66.0	54.3	79.4
	5dB	76.9	54.8	42.4	65.2
	10dB	57.1	44.0	34.7	50.2
	15dB	41.6	36.0	29.8	38.5
BAB	0dB	79.7	66.2	54.5	73.2
	5dB	67.5	53.3	43.9	63.7
	10dB	51.5	44.1	36.0	53.8
	15dB	40.3	36.2	31.35	44.9
HF	0dB	81.2	60.0	46.8	74.6
	5dB	65.2	50.1	38.0	61.7
	10dB	48.6	41.6	32.2	48.9
	15dB	37.4	34.2	29.0	39.4
Average		61.2	48.8	39.4	57.7

Table 1. Phoneme Error Rate (in %) for VTS-AM and DAE for TIMIT dataset

	SNR	No Enhancement	VTS-AM	DAE seen condition	DAE unseen condition
F-16	0dB	89.1	76.1	55.3	88.1
	5dB	70.7	51.5	33.8	67.1
	10dB	41.6	31.7	23.2	38.7
	15dB	23.3	21.2	18.7	24.5
BAB	0dB	85.0	80.3	58.7	91.1
	5dB	59.7	54.8	35.9	73.0
	10dB	34.2	32.1	23.6	52.1
	15dB	21.9	21.0	18.9	40.0
HF	0dB	83.7	70.6	45.9	86.5
	5dB	61.9	48.1	30.1	66.9
	10dB	35.8	31.6	22.3	47.8
	15dB	22.9	21.9	18.5	36.2
Average		52.5	45.1	32.1	59.3

Table 2. Word Error Rate (in %) for VTS-AM and DAE for Librispeech

Initially, we tried to stack the VTS-AM technique and the DAE based enhancement by training the DAE on the VTS-AM enhanced features. However, it did not result in significant performance gain. We also tried to stack the two techniques the reverse way, i.e. by using VTS-AM enhancement on the DAE enhanced features. This also did not result in any gain. Therefore, it was concluded that the two methods cannot be used on top of each other. However, if from the DAE output, it can be determined that the DAE has failed to work, then we can fall back on VTS-AM. It was hypothesized in Section 4, that the SNR of the DAE enhanced feature will be lower in the unseen conditions than the seen conditions. Table 3 shows SNR values computed according to the method proposed in Section 4 for features enhanced by different methods. It can be observed that SNR values of the DAE output for unseen condition is lower as we had hypothesized. The SNR values for other methods (VTS-AM output and DAE seen output) are also provided for comparison and it can be clearly seen that their SNR values are higher. Based on these SNR values, a threshold of 85dB served as an excellent indicator as to whether the DAE has failed to enhance the features. Notice that SNRs of even 75dB are being designated as low in this section. This is because these SNRs are not true SNRs. They are low in the sense that they are lower than the SNR for other methods.

		TIMIT			Librispeech		
		VTS_AM	DAE seen condition	DAE unseen condition	VTS_AM	DAE seen condition	DAE unseen condition
F-16	0dB	133.56	123.30	79.20	148.23	159.91	74.90
	5dB	137.48	122.59	75.40	162.57	165.02	71.99
	10dB	144.49	121.10	72.17	174.04	170.42	72.48
	15dB	149.88	118.84	70.83	183.95	175.02	76.63
BAB	0dB	140.57	119.31	59.05	150.70	170.24	34.23
	5dB	144.99	116.10	61.90	166.74	174.56	41.07
	10dB	151.84	114.85	66.93	181.86	177.66	47.20
	15dB	156.32	113.95	72.72	193.68	178.83	51.80
HF	0dB	139.16	129.04	51.11	167.46	181.86	2.38
	5dB	146.56	128.60	52.59	179.03	180.22	5.17
	10dB	152.03	127.00	59.70	189.04	180.30	11.86
	15dB	155.40	125.37	71.42	197.56	181.62	21.62
Average		146.02	121.67	66.08	174.57	174.63	42.61

Table 3. SNR values (in dB) computed as per Section 4 after different enhancement techniques on TIMIT and Librispeech

In Table 4, we can observe the efficacy of the proposed scheme. On both TIMIT and Librispeech databases and in all conditions, a significant gain in performance can be seen. The PER in unseen conditions improved from 57.79% to 49.34% in case of TIMIT, while the WER for unseen conditions improved from 59.39% to 45.14% in case of Librispeech.

		TIMIT		Librispeech	
	SNR	Integrated seen condition	Integrated unseen condition	Integrated seen condition	Integrated unseen condition
F-16	0dB	54.3	66.2	55.9	76.1
	5dB	42.5	55.6	34.6	51.5
	10dB	34.9	44.5	23.5	31.7
	15dB	30.3	36.8	19.0	21.2
BAB	0dB	54.7	66.1	58.9	80.3
	5dB	44.1	53.2	36.4	54.8
	10dB	36.3	44.4	24.3	32.1
	15dB	31.3	37.4	19.3	21.0
HF	0dB	47.0	60.1	46.7	70.6
	5dB	38.1	50.1	30.6	48.1
	10dB	32.4	41.7	22.6	31.6
	15dB	29.2	36.0	18.7	22.1
Average		39.6	49.3	32.5	45.1

Table 4. Phoneme Error Rate (in %) for TIMIT and Word Error Rate (in %) for Librispeech obtained using the proposed front-end processing (Comparison with results in Table 1 and Table 2 show relative improvement)

8. CONCLUSION

In this paper, we proposed a robust integrated approach for speech recognition in noisy conditions. We have shown that while TDNN based DAE provides significant performance gain in seen conditions, it does not perform as well in unseen conditions. We have also shown that whether the DAE has failed to enhance a signal can be known from the SNR of the

ISBN 978-90-827970-1-5 © EURASIP 2018

DAE enhanced signal. A new approach to estimate the SNR of the DAE enhanced signal has been described and a new approach to integrate VTS-AM and DAE technique has been proposed. The integrated approach performs well in both seen and unseen conditions. We are currently studying the effect of frame selection and root compression on this integrated approach.

9. REFERENCES

[1] H. K. Maganti and M. Matassoni, "An auditory based modulation spectral feature for reverberant speech recognition," *in Proc. INTERSPEECH*, 2010.

[2] O. Vinyals and S. V. Ravuri, "Comparing multilayer perceptron to deep belief network tandem features for robust asr," *in Proc. ICASSP*, pp. 4596–4599, 2011.

[3] N. Moritz, J. Anemller, and B. Kollmeier, "An auditory inspired amplitude modulation filter bank for robust feature extraction in automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 1926–1937, Nov 2015.

[4] B. Das and A. Panda, "Robust front-end processing for speech recognition in noisy conditions," *in Proc. ICASSP*, pp. 5235–5239, March 2017.

[5] 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 Proc. INTERSPEECH*, pp. 895–899, 2014.

[6] 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," *in Proc. IEEE ASRU*, pp. 65–70, Dec 2007.

[7] B. Das and A. Panda, "Psychoacoustic model compensation for robust continuous speech recognition in additive noise," *in Proc. ISSPIT*, pp. 511–515, Dec 2015.

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

[9] B. Li and K. C. Sim, "Noise adaptive front-end normalization based on vector taylor series for deep neural networks in robust speech recognition," *in Proc. ICASSP*, pp. 7408–7412, May 2013.

[10] B. Das and A. Panda, "Vector taylor series expansion with auditory masking for noise robust speech recognition," *in Proc. ISCSLP*, 2016.

[11] Y. Qian, M. Bi, T. Tan, and K. Yu, "Very deep convolutional neural networks for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2263–2276, Dec 2016.

[12] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," *in Proc. INTERSPEECH*, pp. 338–342, 2014.

[13] X. Feng, Y. Zhang, and J. Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," *in Proc. ICASSP*, pp. 1759–1763, May 2014.

[14] Andrew L Maas, Quoc V Le, Tyler M O'Neil, Oriol Vinyals, Patrick Nguyen, and Andrew Y Ng, "Recurrent neural networks for noise reduction in robust asr," *in Proc. INTERSPEECH*, p. 2225, September 2012.

[15] E. Marchi, F. Vesperini, F. Eyben, S. Squartini, and B. Schuller, "A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional lstm neural networks," *in Proc. ICASSP*, pp. 1996–2000, April 2015.

[16] Cong-Thanh Do and Yannis Stylianou, "Improved automatic speech recognition using subband temporal envelope features and time-delay neural network denoising autoencoder," *in Proc. INTERSPEECH*, 2017.

[17] Alexander Waibel, Toshiyuki Hanazawa, Geofrey Hinton, Kiyohiro Shikano, and Kevin J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, pp. 393–404, 1990.

[18] 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.

[19] 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 Proc. IEEE ASRU*, Dec. 2011.

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

ISBN 978-90-827970-1-5 © EURASIP 2018