ISSN: 0976-3031

Available Online at http://www.recentscientific.com

CODEN: IJRSFP (USA)

International Journal of Recent Scientific Research
Vol. 9, Issue, 1(J), pp. 23541-23544, January, 2018

International Journal of Recent Scientific Research

DOI: 10.24327/IJRSR

Research Article

IMPLEMENTATION OF TEST NORMALIZATION TO IMPROVE THE ROBUSTNESS OF AN AUTOMATIC SPEECH RECOGNITION SYSTEM

Gogoi, Swapnanil*

Gauhati University Institute of Distance and Open Learning, Guwahati 781014, Assam, India

DOI: http://dx.doi.org/10.24327/ijrsr.2018.0901.1488

ARTICLE INFO

Article History:

Received 15th October, 2017
Received in revised form 25th October, 2017
Accepted 23rd December, 2017
Published online 28th January, 2018

Key Words:

Automatic Speech Recognition, Robustness, Test Normalization, Sub-band Spectral Subtraction, Hidden Markov Model.

ABSTRACT

This paper presents a test normalization (TN) technique to improve the robustness of an Automatic Speech Recognition (ASR) system. Earlier experimentations of automatic speech recognition show that the application of sub-band spectral subtraction (SSS) is very useful for the reduction of the effect of noise from speech at signal level to improve the recognition accuracy rate in different testing and training conditions. In this work, at the testing process, the estimated scores of each speech signal are normalized with a test normalization technique. Hidden Markov Model (HMM) is used for training and testing process of the ASR system. The ASR experimentation result shows a little improvement of robustness with the combination of SSS and TN approach.

INTRODUCTION

Maintaining the recognition accuracy by an Automatic Speech Recognition (ASR) system in case of different testing and training environments is called the robustness of the system. The recognition accuracy of ASR systems is always affected by the presence of noise in the speech signals. So improvement of the performance of an ASR system is always depending upon the improvement of robustness of the system. It has been observed that different approaches are available to improve the robustness. These techniques have been applied at three different levels that are signal level, feature level and model level. It has been observed that at signal level, spectral subtraction (Martin, 1994; Matsumoto *et al*, 1996; Junhui *et al*, 2002; Karam *et al*, 2014; Upadhyay *et al*, 2015; Gogoi *et al*, 2016) is a popular method to reduce the noise from a noisy spectrum. At feature level, different feature selection (Koniaris *et al*, 2010), feature recombination (Okawa *et al*, 1998), feature compensation (Cui *et al*, 2005) and feature normalization (Viikki *et al*, 1998; Kumar, 2015; Rehr *et al*, 2015) approaches can be applied to improve the robustness. Different types of noise adaptive training strategies and model adaptation approaches (Jiang *et al*, 2000; Nasersharif *et al*, 2005; Xu *et al*, 2007; Kalinli *et al*, 2010; Ganitkevitch, 2015) are also available that can be applied at the model level for the improvement of robustness.

Main objective of this research work is to find out the improvement of robustness of the ASR system due to the implementation of test normalization technique along with the implementation of sub-band spectral subtraction.

METHODOLOGY

In the present work, Hidden Markov Model has been applied to develop the ASR system for the recognition of 12 English vowel sounds embedded in12 h-V-d syllables. Noisy speech data sets have been prepared by adding different types of noises into the clean speech signals. MFCCs (Mel Frequency Cepstral Co-efficient) have been extracted from each speech signals and used as speech feature set for the speech recognition purpose.

In the speech enhancement part, SSS has been used to minimize noise from noisy speech signals. In the final part, TN approach has been applied to improve the robustness of the ASR system.

Techniques Used to Improve Robustness of the ASR System

In the present work, two techniques have been used to improve robustness of the ASR system that are sub-band spectral subtraction and test normalization. SSS has been used to

Corresponding author:* **Gogoi, Swapnanil
Gauhati University Institute of Distance and Open Learning, Guwahati 781014, Assam, India

minimize noise from each of the speech signals and TN has been used to normalize the estimated HMM scores of a particular testing speech signal related to 12 English vowel sounds.

Sub-band Spectral Subtraction (SSS)

Spectral subtraction (SS) is a noise reduction technique where the noise spectrum is estimated form the speech contaminated with noise and then it is subtracted from the spectrum of the speech (Apte, 2012). In the present work, a Sub-band Spectral Subtraction (SSS) approach has been applied as speech enhancement technique at signal level. In SSS approach, the noisy speech has been decomposed into multiple sub band signals and noise is reduced from each sub band signals using SS based on minimum statistics (Martin, 1994). In the final step, noise reduced sub band signals are combined to achieve the original speech signal. The complete process of SSS (Gogoi *et al*, 2016) performed in this work is stated as follows:

- A convolution operation has been performed between the noisy speech and a window based low pass FIR filter to achieve the low-band signal from the noisy speech. The window based lowpass FIR filter has been designed with filter order 24 and frequency constrained 0.5.
- Noise has been reduced from the low-band signal by applying spectral subtraction technique based on minimum statistics.
- A convolution operation has been performed between the noisy speech and a window based high pass FIR filter to achieve the high-band signal from the noisy speech. The Window based high pass FIR filter has also been designed with filter order 24 and frequency constrained 0.5.
- Noise has been reduced from the high-band signal by applying spectral subtraction method based on minimum statistics.
- Finally, both noise reduced low-band and high-band signals have been combined to obtain the noise reduced original speech signal.

Test Normalization

In the recognition process using HMM training data, for each test utterance, 12 probability scores are estimated related to 12 English vowel phonemes (/i/, /ɪ/, /e/, /ɛ/, /æ/, /a/, /ɔ/, /o/, /U/, /u/, /ʌ/,/ɝ/) using the HMM training parameters and the recognition result will be the phoneme whose relative probability score is the highest among all the estimated scores. A test normalization approach has been applied in this present work to normalize the estimated HMM scores of a particular speech signal (figure 1) so that the recognition performance of the ASR system can be improved. This approach is stated as follows:

Mean of all estimated HMM scores except the score related to phoneme m (/i/, /ɪ/, /e/, /ɛ/, /æ/, /a/, /ɔ/, /o/, /U/, /u/, /ʌ/,/ɝ/), μ_m and standard deviation of all estimated HMM scores except the score related to phoneme m, σ_m are estimated as shown in eq. 1 and eq. 2 and $S_o(p)$ is the original estimated HMM score related to phoneme p .

$$\mu_m = \frac{1}{11}\sum_{p\,!=\,m} S_o(p) \tag{1}$$

$$\sigma_m = \sqrt{\frac{1}{11}\sum_{p!=m}(S_o(p)-\mu_m)^2} \tag{2}$$

Normalized score for phoneme m has been estimated by using eq. 3.

$$\mathrm{Norm}_m = S_o(m) - \frac{\mu_m}{\sigma_m} \tag{3}$$

Finally, recognition decision has been performed with these 12 normalized HMM scores (Norm_m) for a particular testing speech signal.

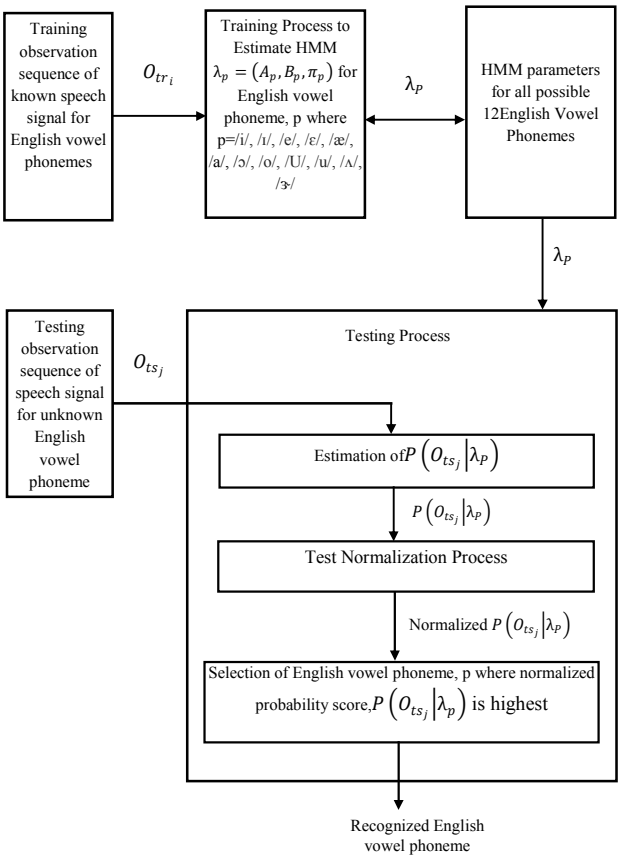

Figure 1 Block Diagram of HMM Training and Testing Phase with Test Normalization Process

Speech Database Preparation

In the present investigation, speech database developed by J. Hillenbrand *et al*. (Hillenbrand *et al*, 2013) has been applied for ASR experimentation. This speech database is composed of male speech signals recorded by 45 adult male speakers, female speech signals recorded by 48 adult female speakers and children voices recorded by 46 children (10 to 12yearold 27 boys and 19 girls). Each speaker has been recorded for 12 vowel sounds/i/, /ɪ/, /e/, /ɛ/, /æ/, /a/, /ɔ/, /o/, /U/, /u/, /ʌ/,/ɝ/embedded in12 h-V-d syllables("heed", "hid", "hayed", "head", "had", "hod", "hawed", "hoed", "hood", "who'd", "hud", "heard", "hoyed", "hide", "hewed" and "how'd"). The speech signals were recorded by a Sony digital audio recorder, Sony PCM-F1 and a microphone, Shure 570-S. Every recorded speech was low-pass filtered at 7.2 kHz where sampling rate

was 16 kHz with 12 bits of amplitude resolution (Hillenbrand *et al*, 1995).

In the present study, the training set has been constructed using first 276 speech signals of adult male speakers and first 288 speech signals of adult female speakers. The test set has been developed by the last 264 adult male speech signals and last 288 adult female speech signals.

For ASR experimentations, to check the recognition accuracy of the ASR system in noisy environment, a noisy speech database has been also developed. It has been performed by artificially contaminating approximately noise free speech signals by 7 different noises (Babble noise, Pink noise, White noise, Volvo noise, Factory noise, Destroyer noise from engine room (Destroyerengine) and Destroyer noise from operations room (Destroyerops)) at 10 dB signal-to-noise ratio (SNR) level. These seven types of noises have been taken from NOISEX-92 database.

Speech Feature Extraction

In the present work, 13 dimensional Mel-Frequency Cepstral Coefficient (MFCC)s have been extracted from each of the training and testing speech signals. In this process, at first pre-emphasis of the speech signal has been performed as shown in eq. 4 for flattening the magnitude spectrum and balancing the high and low frequency components (Loweimi *et al*, 2013).

$$S_p(n) = S(n) - A_p S(n-1) \quad (4)$$

In the present work, it has been considered that $A_p = 0.95$.

After pre-emphasis, each speech signal has been segmented into multiple frames with frame size 25 ms, frame shifting size 10 ms and frame frequency 100 Hz.

After framing process, Hamming window has been applied to each speech frame for smoothing the speech signal as shown in eq. (6) (Rabiner *et al*, 2009).

$$H_m(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1 \quad (5)$$
$$= 0 \qquad otherwise$$

$$S_H(n) = \sum_{n=1}^{N} S_{fr}(n) H_m(n) \quad (6)$$

In this work, it has been considered that $N = 512$.

Then speech frames have been transformed into frequency domain from time domain using Discrete Fourier Transformation (DFT) where DFT is performed by Cooley-Tukey Fast Fourier Transformation (FFT) algorithm. In the next step, speech signals have been represented in Mel-scale by using a triangular bandpass filter bank. Eq. (7) has been applied for this purpose where F_{mel} is the Mel frequency of the linear frequency F. In this work, 20 filters Mel-filter bank has been applied.

$$F_{mel} = 2595\, log_{10}\left(1 + \frac{F}{700}\right) \quad (7)$$

Discrete Cosine Transformation (DCT) has been performed on the logarithm of the mel-power signal spectrum to compute the 13 dimensional MFCC features. To minimize the dissimilarity between the lower order and higher order MFCCs, estimated MFCCs have been weighted by a Lifter Weighting (LW) function as shown in eq.8 and eq. 9 (Juang *et al*, 1987).

$$L_w(k) = 1 + \frac{L}{2}\sin\left(\frac{\pi k}{L}\right), \quad 1 \le k \le L \quad (8)$$
$$= 0 \qquad otherwise$$

$$M_L(n) = M(n) L_w, \quad n = 1,2,3, \ldots\ldots, D \quad (9)$$

In the present work, it has been considered that $L = 22$ and $D = 13$.

Finally, the first and second order derivatives of lifter weighted MFCCs have been computed and combined with the 13-dimensional MFCC to obtain a 39-dimensional speech feature (Arora, 2013; Sharma *et al*, 2014).

RESULTS AND DISCUSSION

For ASR experiment, a HMM with 8 states and 12 Gaussian mixture models has been implemented for training and testing process. Experimental results are represented in table 1. From these results it has been observed that robustness of the ASR system is significantly improved due to SSS (Gogoi *et al*, 2016). Now for further improvement of noise robustness has also been observed by applying Test Normalization (TN) approach but it is not so significant. Even in case of noisy speech signals contaminated by Destroyerengine noise and Destroyerops noise, the recognition accuracy rates are little degraded after application of TN approach.

Table 1 Speech Recognition Rates (in %) with SSS and Combination of SSS and TN

Testing speech database	Without using any speech enhancement technique	With SSS	With SSS and TN
Noise free	90.94	91.12	91.12
With Babble noise	56.88	63.95	64.13
With Pink noise	85.51	87.32	87.68
With White noise	80.98	85.51	85.51
With Volvo noise	87.86	88.59	88.59
With Factory noise	81.34	85.51	85.69
With Destroyerengine noise	53.44	65.22	65.04
With Destroyerops noise	76.63	78.99	78.80

CONCLUSION

Implementation of TN approach has been performed to improve the robustness of ASR system. It has been observed form past experimentations that SSS approach is an efficient technique for the improvement of robustness in case of all types of noisy speech signals used in this research work (Gogoi *et al*, 2016). For further improvement of recognition accuracy, TN has also been found useful in case of noisy speech signals contaminated with Babble noise, Pink noise and Factory noise. But it has been observed that TN approach cannot be able to improve the recognition rate significantly.

References

Apte, D. S. D. (2012): *Speech and Audio Processing*: Wiley India.

Arora, S. V. (2013): Effect of Time Derivatives of MFCC Features on HMM Based Speech Recognition System:

ACEE international Journal on Signal and Image Processing, 4(3): 50–55.

Cui, X. and Alwan, A. (2005): Noise robust speech recognition using feature compensation based on polynomial regression of utterance SNR: *IEEE Transactions on Speech and Audio Processing*, 13(6): 1161-1172.

Ganitkevitch, J. (2015): Speaker adaptation using maximum likelihood linear regression.: *Rheinish-Westflesche TechnischeHochschule Aachen, the course of Automatic Speech Recognition, www-i6. informatik. rwthaachen.* : [Online]. Available: http://www.cs.jhu.edu/~juri/pdf/mllr-rwth-2005.pdf. , Accessed: May 22, 2015.

Gogoi, S. and Bhattacharjee, U. (2016): A Statistical Analysis on the Impact of Speech Enhancement Techniques on the Feature Vectors of Noisy Speech Signals for Speech Recognition: *International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR),* 6(5): 59-66.

Hillenbrand, J. *et al*. (1995): Acoustic characteristics of American English vowels: *The Journal of the Acoustical society of America*, 95(5): 3099-3111.

Hillenbrand, J. *et al*. (2013): Vowel database: [Online]. Available: http://homepages.wmich.edu/~hillenbr/voweldata.html., Accessed: Jun. 9, 2013.

Jiang, H., Hirose, K. and Hue, Q. (2000): A minimax search algorithm for robust continuous speech recognition: *IEEE Transactions on Speech and Audio Processing*, 8(6): 688–694.

Juang, B. H., Rabiner, L. R. and Wilpon, J. G. (1987): On the Use of Bandpass Liftering in Speech Recognition: *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-35(7): 947-954.

Junhui, Z., Xiang, X. and Jingming, K. (2002): Noise suppression based on auditory-like filters for robust speech recognition: in *6th International Conference on In Signal Processing*, IEEE, pp. 560-563.

Kalinli, O. *et al*. (2010): Noise adaptive training for robust automatic speech recognition: *IEEE Transactions on Audio, Speech, and Language Processing*, 18(8): 1889-1901.

Karam, M. *et al*. (2014): Noise removal in speech processing using spectral subtraction: *Journal of Signal and Information Processing*.

Koniaris, C., Kuropatwinski, M. and Kleijn, W. B. (2010): Auditory-model based robust feature selection for speech recognition: *The Journal of the Acoustical Society of America*, 127(2): EL73-EL79.

Kumar, D. S. (2015): Feature Normalisation for Robust Speech Recognition: arXiv preprint arXiv: 1507.04019.: Available: https://arxiv.org/pdf/1507.04019 .pdf., Accessed: Dec. 13, 2015.

Loweimi, E. *et al*. (2013): On the Importance of Pre-emphasis and Window Shape in Phase-Based Speech Recognition: in *International Conference on Nonlinear Speech Processing*, Berlin: Springer, pp.160-167.

Martin, R. (1994): Spectral subtraction based on minimum statistics: *Power*, 6(8).

Matsumoto, H. and Naitoh, N. (1996): Smoothed spectral subtraction for a frequency-weighted HMM in noisy speech recognition: in *Fourth International Conference on In Spoken Language*, IEEE, pp. 905-908.

Nasersharif, B. and Akbari, A.(2005):Improved HMM entropy for robust sub-band speech recognition: in*13th European In Signal Processing Conference*, IEEE, pp. 1–4.

NOISEX92 noise database: [Online]. Available: http:// spib.rice.edu/spib/ select noise. html., Accessed: Dec. 20, 2013.

Okawa, S., Bocchieri, E. and Potamianos, A. (1998):Multi-band speech recognition in noisy environments: in *IEEE International Conference on In Acoustics, Speech and Signal Processing*, IEEE, pp. 641-644.

Rabiner, L. R. and Schafer, R. W. (2009): *Digital processing of speech signals*: Pearson Education.

Rehr, R. and Gerkmann, T. (2015): Cepstral noise subtraction for robust automatic speech recognition: in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 375-378.

Sharma, S., Shukla, A. and Mishra, P. (2014): Speech and Language Recognition using MFCC and DELTA-MFCC: *International Journal of Engineering Trends and Technology (IJETT)*, 12(9): 449-452.

Upadhyay, N. and Karmakar, A.(2015): Speech Enhancement using Spectral Subtraction-type Algorithms: A Comparison and Simulation Study: *Procedia Computer Science*, 54: 574-584.

Viikki, O. and Laurila, K.(1998): Cepstral domain segmental feature vector normalization for noise robust speech recognition: *Speech Communication*, 25(1), 133-147.

Xu, H. *et al*. (2007): Noise Condition-Dependent Training Based on Noise Classification and SNR Estimation: *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8): 2431–2443.

How to cite this article:

Gogoi, Swapnanil.2018, Implementation of Test Normalization To Improve The Robustness of An Automatic Speech Recognition System. *Int J Recent Sci Res.* 9(1), pp. 23541-23544. DOI: http://dx.doi.org/10.24327/ijrsr.2018.0901.1488

****