Vocal Tract Length Normalization and Sub-Band Spectral Subtraction Based Robust Assamese Vowel Recognition System

Swapnanil Gogoi
GUIDOL
Gauhati University
Guwahati, Assam, India, Pin – 781 014
swapnanil22@gmail.com

Utpal Bhattacharjee
Department of Computer Science and Engineering
Rajiv Gandhi University
Doimukh, Arunachal Pradesh, India, Pin -791 112
utpal.bhattacharjee@rgu.ac.in

***Abstract* -In this paper, vocal tract length normalization (VTLN) and sub-band spectral subtraction (SSS) have been used for speaker adaptation and noise reduction to develop an Assamese vowel recognition system which is robust to the speaker and environment variabilities. In the present work VTLN has been implemented to reduce the effects of inter speaker variabilities and sub-band spectral subtraction has been used to reduce the effects of environmental variabilities. The effectiveness of VTLN in noisy and noise-free environment has been evaluated for Assamese vowel recognition system. The Assamese vowel recognition system has been implemented using Hidden Markov Model (HMM). Mel Frequency Cepstral Coefficient (MFCC) has been used as feature vector. Experimented result shows that the performance of the system improved considerably after applying VTLN technique in noise-free and some of the noisy conditions.**

Keywords - Automatic Speech Recognition,Vocal Tract Length Normalization, Sub-band Spectral Subtraction, Hidden Markov Model.

I. INTRODUCTION

The performance of automated speech recognition (ASR) system is degraded in case of mismatched training and testing conditions [1]. Different approaches have been investigated in past to reduce the noise from speech signals like Spectral subtraction, Wiener filtering, Kalman filtering etc [2, 3, 4, 5, 6]. Inter speaker variations is another reason for the performance degradation of ASR systems. Physiological variations among different speakers are one of the main cause of inter speaker variations [1].Different Vocal Tract Lengths (VTL) among different speakers is one of the major physiological source for the induction of inter speaker acoustic variations. In this context, it is observed that VTL can vary from approximately 13 cm for adult females to over 18 cm for adult males [7, 8]. VTLN has been found as an effective speaker normalization approach in some past research works to reduce this type of inter speaker variations from speech signals [9, 10, 11, 12].

The main objective of the present work is to investigate the improvement of recognition performance due to the implementation of VTLN in both noise free and noisy conditions in case of an Assamese vowel recognition system.

A HMM based Assamese vowel recognition system has been developed in the present work. The training and testing speech signals are recorded in approximately noise free environment. A noisy testing speech database has also been constructed by adding different noises to the noise free testing speech signals. Sub-band Spectral Subtraction approach has been used to minimize the effect of noise from the noisy speech signals. To introduce inter speaker variations, training process has been performed with only male speech signals and testing process has been performed with only female speech signals and vice versa. Finally VTLN has been implemented to reduce the inter speaker variations from training and testing speech signals.

II. SPEECH DATABASE PREPARATION

The Assamese is the main language in Assam, north-eastern state of India. In Assamese language, thirty two essential phonemes are available where total number of vowel phonemes is 8. The vowels are ই (/i /) , এ (/ɛ/) , এ' (/e /) , আ (/a/) , অ' (/ɔ/) , অ (/ɒ/) , ও(/o/) and উ (/u/) [13].

In this research work, a speech database is prepared with 10 adult male and 10 adult female speakers to perform the ASR experiments. The speakers belong to the age group from 25 years to 45 years and each speaker is recorded five times for each vowel phoneme. Recording has been done at 16 kHz sampling rate mono-channel and at 16 bits resolution. Recording has been performed in a controlled and approximately noise free acoustical environment.

From the Assamese vowel database, two sets have been prepared. The first set has been used for training and the second set has been used for testing. The training set consist of speech signal from 5 male and 5 female speakers and the testing

set consist of the speech from remaining speakers. A noisy version of the testing speech database has been prepared by adding seven different types of noises of NOISEX-92 [14] database at 10dB SNR to each speech signal of the noise free testing speech set. The noises considered in the present study are, Babble noise, Pink noise, White noise, Volvo noise, Factory noise, Destroyer noise from engine room (Destroyerengine) and destroyer noise from operations room (Destroyerops).

III. FEATURE EXTRACTION

Mel-Frequency Cepstral Coefficient (MFCC) has been extracted from each speech signals. The speech signal is segmented into 25 msec frame with frame rate 100 Hz. Hamming window has been applied for smoothing the speech signal. A pre-emphasis filter $H(z)=1-0.96z^{-1}$ has been applied before framing. The pre-emphasized speech signal is segmented into frame of 20 microseconds with frame frequency 100 Hz. Each frame is multiplied by a Hamming window. From the windowed frame, FFT has been computed and the magnitude spectrum is filtered with a bank of 21 triangular filters spaced on Mel-scale and constrained into a frequency band of 300-3400 Hz. The log-compressed filter outputs are converted to cepstral coefficients by Discrete Cosine Transformation. The 0th cepstral coefficient is not used in the cepstral feature vector since it corresponds to the energy of the whole frame, and only first 13 MFCC coefficients have been used. Then to capture the dynamic property of the speech signal, the first and second order derivatives of extracted MFCC are also combined with the 13-dimensional MFCC to achieve a 39-dimensional speech feature [15, 16].

IV. SPEECH ENHANCEMENT TECHNIQUE APPLIED

A. *Sub-band Spectral Subtraction (SSS)*

In this research work, a Sub-band Spectral Subtraction (SSS) technique has been applied to reduce noise from noisy speech signals. In this approach, band-pass filter has been used to separate different frequency bands of the speech signal. Noise has been estimated for each frequency band separately. Then noise is reduced from each sub-band using spectral subtraction based on minimum statistics [17]. Finally, the original speech signal has been computed from the noise reduced sub-band signals.

B. *Vocal Tract Length Normalization (VTLN):*

VTLN has been implemented in the present work by warping the frequency axis of the power spectrum. The selection of proper warping factor for each speaker has been performed by a grid search approach from a set of 13 possible warping factors (ξ) from 0.88 to 1.12 with step size 0.02[7,8,9]. A piecewise linear warping function has been applied for frequency warping (eq. (1)).

$$F_w = \xi F \quad (1)$$

Where F_w is the warped frequency and F is the input frequency.

VTLN has been implemented to normalize both the training and testing speech signals by frequency warping and in this process eq. (2) [8, 18] has been used as warping factor selection criterion that is based on maximum likelihood.

$$\hat{\alpha} = \text{argmax} P_r(T^{\alpha}|\lambda, W) \quad (2)$$

Where

- $\hat{\alpha}$: Optimal warping factor
- λ : The set of HMM models
- W : Utterance
- T^{α} : The acoustic observation vector computed with the warping factor α

V. EXPERIMENTS AND RESULTS

Initially, experiments have been performed for Automatic Speech Recognition with both noise free and noisy speech signals without implementing SSS and VTLN. From these experiments it has been observed that due to presence of noise the recognition performance has been degraded. In case of noise free speech signals the average recognition rate has been observed as 83.04% (table 1).

In the second step of the experimentation, VTLN has been implemented on only noise free speech signals and in this case it has been observed that average recognition accuracy is increased by 3.57% that is 86.61% (table 1).

In the third step of experimentation SSS has been performed for noise reduction from the noisy speech signals and recognition rates in case of different noisy conditions are shown in table 2. From these results, it has been observed that recognition accuracy has been increased due to speech enhancement by SSS in case of all noisy conditions except in case of Volvo and Factory noise.

TABLE 1: ASSAMESE VOWEL RECOGNITION RATES (IN %) IN NOISE FREE CONDITION USING VTLN

Training	Testing	Without VTLN	With VTLN
Male	Female	81.35	87.6
Female	Male	84.73	85.62

TABLE 2: ASSAMESE VOWEL RECOGNITION RATES (IN %) IN DIFFERENT NOISY CONDITIONS USING SSS

Noise Type	Without Speech Enhancement	Using SSS
Babble	72.32	74.11
Pink	78.57	82.14
White	68.75	73.21
Volvo	78.57	78.57
Factory	77.68	76.79
Destroyerengine	52.68	67.86
Destroyerops	57.14	76.79

TABLE 3: ASSAMESE VOWEL RECOGNITION RATES (IN %) IN DIFFERENT NOISY CONDITIONS USING VTLN AND COMBINATION OF SSS AND VTLN

Noise Type	Using VTLN	Using SSS and VTLN
Babble	72.32	69.64
Pink	72.32	79.46
White	61.61	75.89
Volvo	79.46	83.93
Factory	75	81.25
Destroyerengine	55.36	69.64
Destroyerops	53.57	70.54

Finally, VTLN has been implemented and recognition rates in different noisy conditions are presented in table 3. From this set of results it has been observed that except in case of Volvo noise, VTLN has not been found as an effective speech enhancement approach in noisy conditions. So VTLN has been implemented on the noise reduced speech signals that are achieved by implementing SSS and recognition results are shown in table 3. From these results it has been observed that combination of SSS and VTLN is also not found to be effective approach in case of some of the noisy conditions. It has been found effective in case of White, Volvo, Factory and Destroyerengine noise.

In the present work, it has also been observed that in most of the cases, 0.88 has been come out as optimal warping factor and three other warping factors that are 0.96, 0.98 and 1.02 also observed to be optimal warping factors in remaining cases.

VI. CONCLUSION

In this research work, a robust Assamese vowel recognition system has been developed to recognize Assamese vowels in mismatched testing and training conditions. The training process has been performed with approximately noise free speech signals and the testing process has been performed with both noise free and noisy speech signals. The noisy versions have been obtained by contaminated the speech signal with seven 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)). SSS approach has been used to reduce the noise from noisy speech signals. On the other hand inter speaker variation has also been introduced by performing training process with male speech signals and testing process with female speech signals and vice versa. VTLN has been applied to reduce the effect of inter-speaker variations and to improve recognition rate. From the ASR experimented results, it has been observed that VTLN is an effective approach to reduce the effect of inter speaker variations so that the recognition accuracy can be enhanced in case of noise free speech signals. On the other hand in case of different noisy conditions it has not been able to improve the robustness of the ASR system. But combination of SSS and VTLN can be useful in case of some of the noisy conditions. In the present work, one drawback of VTLN has also been observed. VTLN require a large amount of time due to the grid search approach for the selection of proper warping factor. So in such situation if the grid search approach can be replaced by some other effective approach then VTLN will be more useful. It has also been observed that noise can be effectively reduced by SSS so that recognition accuracy can be improved.

References:

[1] L. Rabiner and B.H Juang, Fundamental of Speech Recognition, New Delhi: Dorling Kindersley (India) Pvt. Ltd, 1993.

[2] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-27, no. 2, pp. 113–120, 1979.

[3] A. R. Fukane and L. S. Shashikant, "Different approaches of spectral subtraction method for enhancing the speech signal in noisy environments," International Journal of Scientific & Engineering Research, vol. 2, no. 5, 2011.

[4] M. Karam, H. F. Khazaal, H. Aglan and C. Cole, "Noise removal in speech processing using spectral subtraction", Journal of Signal and Information Processing, 2014.

[5] A. Agarwal and Y. M. Cheng, "Two-stage Mel-warped Wiener filter for robust speech recognition", In Proc. ASRU, vol. 99, pp. 67-70, 1999.

[6] G. R. Babu and R. Rao, "Modified Kalman Filter-based Approach in Comparison with Traditional Speech Enhancement Algorithms from Adverse Noisy Environments", International Journal on Computer Science and Engineering, vol. 3, no. 2, pp. 744-759, 2011.

[7] D. Giuliani, M. Gerosa, and F. Brugnara, "Improved automatic speech recognition through speaker normalization," Computer Speech & Language, vol. 20, no. 1, pp. 107–123, 2006.

[8] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in IEEE International Conference on Acoustics, Speech, and Signal Processing, IEEE, vol. 1, 1996.

[9] J. Lung et al., "Implementation of Vocal Tract Length Normalization for Phoneme Recognition on TIMIT Speech Corpus," in International Conference on Information Communication and Management, Singapore: IPCSIT, pp. 136–140, 2011.

[10] B . Widmer, "Implementation of Vocal Tract Length Normalization A Study of Methods". [Online]. Available: http : // ssli.ee.washington.edu / people / bwidmer /VTL_Talk/VTL_Talk.PDF. Accessed: 2013.

[11] S. Gogoi and U. Bhattacharjee, "Impact of Vocal Tract Length Normalization on the Speech Recognition Performance of an English Vowel Phoneme Recognizer for the Recognition of Children Voices," International Journal of Computer Trends and Technology (IJCTT), vol. 39, no. 2, pp. 105–109, 2016. [Online]. Available: http://www.ijcttjournal.org/2016/Volume39/number-2/IJCTT-V39P118.pdf. Accessed: Oct. 1, 2016.

[12] G. Garau, S. Renals, and T. Hain, "Applying Vocal Tract Length Normalization to Meeting Recordings," 2005. [Online].Available: http://www.cstr.ed.ac.uk/downloads/ publications/2005/giuliagarau_eurospeech05.pdf. Accessed: May 8, 2013.

[13] B. Kakati, Assamese, its Formation and Development, 5th ed. Guwahati: LBS Publications, 2007.

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

[15] S. Sharma, A. Shukla, and P. Mishra, "Speech and Language Recognition using MFCC and DELTA-MFCC," *International Journal of Engineering Trends and Technology (IJETT)*, vol. 12, no. 9, pp. 449–452, 2014.

[16] S. V. Arora, "Effect of Time Derivatives of MFCC Features on HMM Based Speech Recognition System," *ACEE international Journal on Signal and Image Processing*, vol. 4, no. 3, pp. 50–55, 2013.

[17] R. Martin, "Spectral subtraction based on minimum statistics," Power, vol. 6, no. 8, 1994.

[18] L. Lee and R. C. Rose, "A frequency warping approach to speaker normalization," IEEE Transactions on Speech and audio processing, vol. 6, no. 1, pp. 49–60, 1998.