Impact of Vocal Tract Length Normalization on the Speech Recognition Performance of an English Vowel Phoneme Recognizer for the Recognition of Children Voices

Swapnanil Gogoi[1], Utpal Bhattacharjee[2]

[1]*Assistant professor & GUIDOL & Gauhati University*

Gopinath Bordoloi Nagar, Dist. Kamrup(M),Guwahati-781014, Assam, India

[2]*Professor & Computer Science and Engineering & Rajiv Gandhi University*

Rono Hill, Dist. Papum Pare, Doimukh, Arunachal Pradesh 791112, India

Abstract *— Differences in human vocal tract lengths can cause inter speaker acoustic variability in speech signals spoken by different speakers for the same textual version and due to these variations, the robustness of a speaker independent (SI) speech recognition system is affected. Speaker normalization using vocal tract length normalization (VTLN) is an effective approach to reduce the affect of these types of variability from speech signals. In this paper impact of VTLN approach has been investigated on the speech recognition performance of an English vowel phoneme recognizer with both noise free and noisy speech signals spoken by children. Pattern recognition approach based on Hidden Markov Model (HMM) has been used to develop the English vowel phoneme recognizer. Here training phase of the automatic speech recognition (ASR) system has been performed with speech signals spoken by adult male and female speakers and testing phase is performed by the children speech signals. In this investigation, it has been observed that use of VTLN can effectively improve the robustness of the English vowel phoneme recognizer in both noise free and noisy conditions.*

Keywords *— Automatic speech recognition, speaker independent, vocal tract lengths, vocal tract length normalization, Hidden Markov model.*

I. INTRODUCTION

Speech patterns for a particular phoneme or word or sentence can be dissimilar depending upon different speakers. Now this type of dissimilarity is introduced due to speakers' physiological variations, gender variations and variations in speakers' regional accent. These inter speaker acoustic variations available in speech signals are always responsible for the degradation of recognition performance of SI ASR systems. Now variations in VTLs among different speakers are observed and it is one of the main physiological source of inter speaker acoustic variability. The VTL of the male speakers is greater than the female speakers. VTL can vary from approximately 13 cm for adult females to over 18 cm for adult males [10, 11]. These variations affect spectral formant frequency position which can cause differences in formant frequencies and because of this the performance of a SI ASR system is degraded [8, 10, 11]. It has been also observed that the speech patterns of children voices are more similar to the female voices than the male voices. Speaker normalization and adaptation approaches can be considered as solutions to reduce the effect of these inter speaker acoustic variability from speech signals so that robustness of SI ASR system can be improved. Maximum Likelihood Linear Regression (MLLR), Maximum A Posteriori (MAP) and VTLN are three popular approaches that can be used as answers to the problem stated above. In 1995, C.J. Leggetter and P.C. Woodland [12] investigated speaker adaptation using MLLR in continuous density HMMs to develop a speaker independent (SI) ASR system with better robustness. In 2011, J.W.J. Lung et al. [14] implemented VTLN to reduce the effect of inter-speaker acoustic variability where phoneme recognition was performed on TIMIT corpus. In 2013, B. Das et al. [13] developed an ASR system with acoustic model adaptation techniques like VTLN, MLLR and MAP for aged population in Bengali language.

Main objective of this paper is to present the impact of VTLN as the speaker normalization techniques on the speech recognition performance of an ASR system that is used to recognize English vowel phonemes where training speech signals are recorded by adult male and female speakers and testing speech signals are recorded by children. In this investigation both noise free and noisy speech signals have been used at the testing phase of the ASR system and on the other hand only noise free signals are used at the training phase.

II. SPEECH DATABASE PREPARATION

In this paper, speech database from [6] has been used for ASR experiments where one part of the database is consist of male speech signals recorded

by 45 male speakers, second part is consist of female speech signals recorded by 48 female speakers and third part is consist of speech signals recorded by 46 children (10 to 12 year old 27 boys and 19 girls). Here each speaker recorded speech signals for 12 vowel phonemes /i, ɪ, e, ɛ, æ, a, ɔ, o, U, u, ʌ, ɜ˞/ embedded in h-V-d syllables ("heed", "hid", "hayed", "head", "had", "hod", "hawed", "hoed", "hood", "who'd", "hud", "heard", "hoyed", "hide", "hewed" and "how'd") in noise free environment. In our experiment, for the training phase, the training speech database has been constructed with the adult male and female speech files. On the other hand, the testing speech database has been developed with the children speech signals for recognition process. Now from this testing speech database, 7 more noisy speech databases have been developed by adding 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)) from NOISEX-92[7] database to each noise free testing speech signals. So here for the recognition phase, 8 speech databases have been used where one is noise free speech database and other 7 are noisy speech databases.

III. SPEAKER NORMALIZATION APPLYING VTLN

VTLN is a speaker normalization approach where inter speaker acoustic variability originated from the differences in vocal tract lengths among different speakers can be reduced by warping the frequency axis of the power spectrum. In this research work, VTLN has been implemented only on testing speech signals and it is performed within the process of Mel-frequency cepstral coefficients (MFCC) estimation.

Estimation of MFCC has been divided into the following processes.

- Pre-emphasis: Pre-emphasis of the speech signal is a process performed by one coefficient digital filter termed as pre-emphasis filter as shown in equation (1) for flattening the magnitude spectrum and balancing the high and low frequency components [15,16]. In this work the value of a_{pre} has been considered as -1.0.

$$S_{pre}(n) = S(n) - a_{pre}S(n-1) \qquad (1)$$

- Framing: After pre-emphasis, the speech signal has been divided into multiple frames with frame size 25 ms and 10 ms as frame shifting size.

- Windowing: After framing process, some discontinuity may be introduced in the speech samples of each frame. So, Hamming window function as shown in equation (2) [1] has been applied to each frame using equation (3) as a solution to this problem which will attenuate the speech samples at the beginning and end of the signal. In equation (3) $S_{hw}(n)$ is the Hamming windowed speech frame of the speech frame $S_{fr}(n)$ and N is the frame size. In this work value of N has been considered as 400.

$$W_{hm}(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \qquad (2)$$

$$For\ 0 \leq n \leq N-1$$

$$= 0 \qquad otherwise$$

$$S_{hw}(n) = \sum_{n=1}^{N} S_{fr}(n)W_{hm}(n) \qquad (3)$$

- Discrete Fourier Transformation (DFT): Now the windowed speech frames have been converted into frequency domain from time domain using DFT by Cooley-Tukey Fast Fourier Transformation (FFT) algorithm.

- Mel-scale Warping: After FFT of the speech signal, it is required to filter the signal using triangular bandpass filters so that the signal can be represented in Mel-scale. For this purpose equation (4) [5] has been used where F_{mel} is the Mel frequency of the linear frequency F. In this work, Mel-filter bank with 20 filters have been constructed.

$$F_{mel} = 2595\ log_{10}\left(1 + \frac{F}{700}\right) \qquad (4)$$

- Log and Discrete Cosine Transformation (DCT): Finally, DCT has been performed using equation (5) on the logarithm of the mel-power signal spectrum to estimate the MFCC features. In equation (5) M is the number of triangular bandpass Mel-filters and D is the number of MFCCs. In this research work, D is considered as 12.

$$MFCC(n) = \sum_{k=1}^{M} S(k)cos\left(\frac{\pi n\left(k-\frac{1}{2}\right)}{M}\right) \qquad (5)$$

$where\ \ n = 1,2,3, \dots, D$

- Lifter Weighting: Now estimated MFCCs have been weighted by a lifter weighting function as shown in equation (6) and equation (7) to reduce the variation between the lower order and higher order MFCCs [2]. In equation (6), value of L has been considered in this work as 22.

$$Lifter_w(k) = 1 + \frac{L}{2}\sin\left(\frac{\pi k}{L}\right) \qquad (6)$$
$$For\ \ 1 \le k \le L$$

$$= 0 \qquad otherwise$$

$$MFCC_L(n) = MFCC(n) Lifter_w \qquad (7)$$

$$where \quad n = 1,2,3, \dots \dots, D$$

Now VTLN has been implemented by frequency warping of the speech power spectrum with a warping factor (α) after performing DFT in the process of MFCC estimation as shown in fig. 1 and here piecewise linear warping function as shown in equation (8) has been used for frequency warping where F_{warp} is the warped frequency and F is the input frequency.

$$F_{warp} = \alpha F \qquad (8)$$

So at the end weighted MFCC has been computed from the frequency warped speech data. After that following steps have been performed to compute the speech feature vectors from the estimated weighted MFCC.

- Log energies of each speech frame have been computed by using equation (9).

$$E_n = log \sum_{k=1}^{M} S(k)^2 \qquad (9)$$

- First time derivatives of MFCCs termed as Delta MFCC (DMFCC) have been computed from the estimated MFCCs and log energies of the speech signals using equation (10) [3, 4] where Δc[t] is a DMFCC of t^{th} frame with k = 2.

$$\Delta c[t] = \frac{\sum_{m=1}^{k} m(c[t+m] - c[t-m])}{2\sum_{m=1}^{k} m^2} \qquad (10)$$

- Second time derivatives of MFCC termed as Delta-delta MFCC (DDMFCC) has been computed from the estimated DMFCCs using equation (10).

- Finally 39-dimensional speech feature vectors for each speech has been estimated by combining 12-dimensional MFCCs, 1-dimensional log energy, 13-dimensional DMFCCs and 13-dimensional DDMFCCs.

The main problem of implementing VTLN is that the selection of the proper warping factor for a particular speaker. In this paper warping factor has been selected by a grid search approach from a set of 13 possible warping factors (α) from 0.88 to 1.12 with step size 0.02 so that it will reflect approximately 25% range in vocal tract lengths available in humans[10]. Here equation (11) has been used as warping factor selection criterion based on maximum likelihood [9, 10, 11] where $\hat{\alpha}$ is the optimal warping factor, λ is the set of HMM models, W is the utterance and X^{α} is the speech feature vectors that are computed with the weighted MFCCs estimated from the warped speech data with warping factor α.

$$\hat{\alpha} = \text{argmax}\ \Pr(X^{\alpha}|\lambda, W) \qquad (11)$$

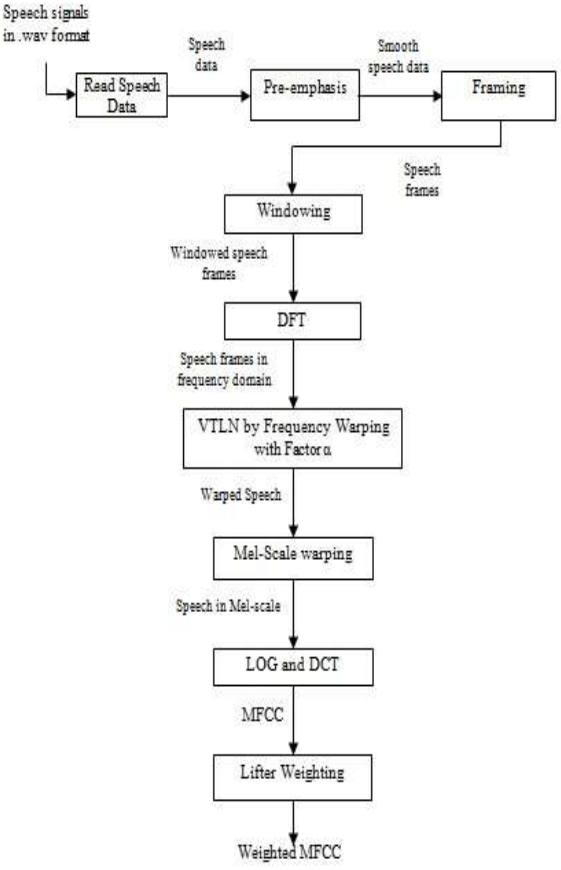

Fig. 1: Block diagram of MFCC estimation process with VTLN

IV. EXPERIMENTS AND DISCUSSION

In this research work, a left to right single Gaussian HMM with 8 states and 12 Gaussian models per state has been implemented for the training and testing process of the English vowel phoneme recognizer.

Initially, ASR experiments have been performed with both noise free and noisy speech signals without applying VTLN. In the next part of the experiments, VTLN has been applied on both noise free and noisy speech signals. The ASR experimented results are shown in table 1. From these results, it has been observed that VTLN can effectively improve the recognition performance in both noise free and noisy conditions. So impact of VTLN on speech recognition performance of the ASR system has been illustrated by the graphical representation of the recognition accuracy improvement rate (in %) correspond to noise free and noisy versions of the speech signals in fig. 2.

Now, in case of VTLN, selection of proper warping factors play an important role in the achievement of the mentioned improved recognition accuracy from the ASR system. In this research work, in most of the cases, the selected warping factor has been observed as 1.12. Among the other possible warping factors, 1.06, 1.08 and 1.10 are the warping factors which are also selected by the grid search approach in some cases.

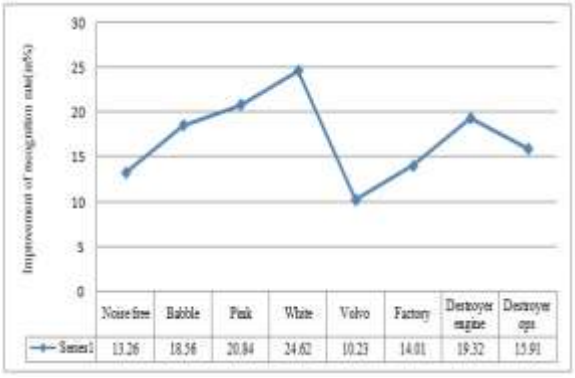

Fig. 2: Improvement of recognition accuracy rate (in %) after application of VTLN correspond to noise free and noisy versions of speech signals

Table 1: Speech Recognition Rate (in %) Before and After Speaker Normalization with VTLN Correspond to Noise Free and Noisy Versions of Testing Speech Signals

Noise type	Without Speaker Normalization	After Speaker Normalization with VTLN
Noise free	75	88.26
Babble	52.65	71.21
Pink	60.98	81.82
White	58.33	82.95
Volvo	75	85.23
Factory	59.47	73.48
Destroyerengine	31.06	50.38
Destroyerops	53.79	69.70

V. CONCLUSIONS

Inter speaker variability is one of the main challenge to develop a robust ASR system and variations in VTL of humans is one of the main sources of inter speaker acoustic variability. So in this research work, VTLN has been implemented to minimize the effect of this variability in the testing phase of the ASR system where the training phase is performed with adult male and female voices and testing phase is performed with children's voices. ASR experiments show great improvement of recognition rate by using VTLN. The main problem of VTLN is to select proper warping factor for each speaker. Here a grid search approach has been applied but it has been observed that this approach is computationally expensive. So if an alternative approach to the grid search technique for selection of proper warping factor can be implemented then VTLN can be more useful. In this research work, 1.12 is the most popular selected warping factor by the grid search approach.

REFERENCES

[1] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals*. Prentice Hall, 1978.

[2] B.H. JUANG, L. R. RABINER, and J. G. WILPON, "On the Use of Bandpass Liftering in Speech Recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-35, no. 7, pp. 947–954, 1987.

[3] 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.

[4] 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.

[5] F. Zheng, G. Zheng, and Z. Song, "Comparison of different implementations of MFCC," *Journal of Computer Science and Technology*, vol. 16, no. 6, pp. 582–589, 2001.

[6] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of American English vowels," *The Journal of the Acoustical society of America*, vol. 95, no. 5, pp. 3099–3111, 1995. [Online]. Available: http://homepages.wmich.edu/~hillenbr/voweldata.html. Accessed: Aug.22,2014.

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

[8] J. Lung and W. Jing, et al., "Implementation of vocal tract length normalization for phoneme recognition on TIMIT speech corpus," in *International Conference on*

Information Communication and Management, IPCSIT, vol. 16, 2011.

[9] 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.

[10] 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.

[11] 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.

[12] C. J. Leggetter and P. C. Woodland. "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models." *Computer Speech & Language, vol.* 9, no. 2, pp. 171-185, 1995.

[13] B. Das, et al. "Aging speech recognition with speaker adaptation techniques: Study on medium vocabulary continuous Bengali speech."*Pattern Recognition Letters*, vol. 34, no. 3 pp. 335-343, 2013.

[14] 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, 2011, pp. 136–140.

[15] J.W. Picone, "Signal modeling techniques in speech recognition," *Proceedings of the IEEE*, vol. 81, no. 9, pp. 1215–1247, 1993.

[16] E. Loweimi, S. M. Ahadi, T. Drugman and S. Loveymi, "On the Importance of Pre-emphasis and Window Shape in Phase-Based Speech Recognition," in *International Conference on Nonlinear Speech Processing*, Berlin: Springer, 2013.