2011 International Conference on Information Communication and Management
IPCSIT vol.16 (2011) © (2011) IACSIT Press, Singapore

Implementation of Vocal Tract Length Normalization for Phoneme Recognition on TIMIT Speech Corpus

Jensen Wong Jing Lung [+], Md.Sah Hj.Salam, Mohd Shafry Mohd Rahim and Abdul Manan Ahmad

Department of Computer Graphics & Multimedia, Faculty of Computer Science and Information System,
University Technology Malaysia, 81310 UTM Skudai, Johor, Malaysia

Abstract. Inter-speaker variability, one of the problems faced in speech recognition system, has caused the performance degradation in recognizing varied speech spoken by different speakers. Vocal Tract Length Normalization (VTLN) method is known to improve the recognition performances by compensating the speech signal using specific warping factor. Experiments are conducted using TIMIT speech corpus and Hidden Markov Model Toolkit (HTK) together with the implementation of VTLN method in order to show improvement in speaker independent phoneme recognition. The results show better recognition performance using Bigram Language Model compared to Unigram Language Model, with Phoneme Error Rate (PER) 28.8% as the best recognition performance for Bigram and PER 38.09% for Unigram. The best warp factor used for normalization in this experiment is 1.40.

Keywords: VTLN, inter-speaker variability, speech signal, warp factor, phoneme recognition.

1. Introduction

Differences in human voices are caused by the different sizes of vocal tract (VT), thus their generated speech signals contain frequencies that are not always constantly the same. The variation in acoustic speech signals from different speakers, adding with different accent, dialect, speaking rate and style, contribute to more problems in matching the trained speech signal accurately in the system. These physiology and linguistic differences between speakers are known to be the inter-speaker variability [8, 11], affecting the overall performance for continuous Automatic Speech Recognition (ASR) system.

One physical source of inter-speaker variability is the vocal tract length (VTL). In Figure 1, this model represents the human vocal apparatus which is the main source of human speech-voice generation. Speech spectrum is shaped by VT that is marked within dotted box, starts with the opening of the vocal cords, or glottis, and ends at the lips and nasal [2]. By using simple analogy on each bottle with different water level generate different frequency, the similarity can be apply that the size and length of VT affects the speech signal's frequency.

Physical difference in VTL is more noticeable between male and female speakers. Male speakers have longer VT that generates lower frequency speech spectrum. On the other hand, female speakers have shorter VT which generates higher frequency speech spectrum. According to Lee & Rose [3,4], VTL can vary from approximately 13 cm for adult females to over 18 cm for adult males. These VTL differences affect the position of spectral formant frequency by as much as 25% between adult speakers. This formant position difference leads to the mismatched formant frequencies, resulting in decreased recognition performance.

Due to these VT differences, speaker independent ASR system that is trained with different speakers, is generally worse than speaker dependent ASR system in recognition performance. ASR modelling efficiency

[+] Corresponding author. Tel.: +60128802126.
E-mail address: jensen_wg@yahoo.com.

is dramatically reduced without the appropriate alignment on the frequency axis from the speech spectrum [12]. Hence, the frequency speech spectrums needed to be approximately scaled linearly.

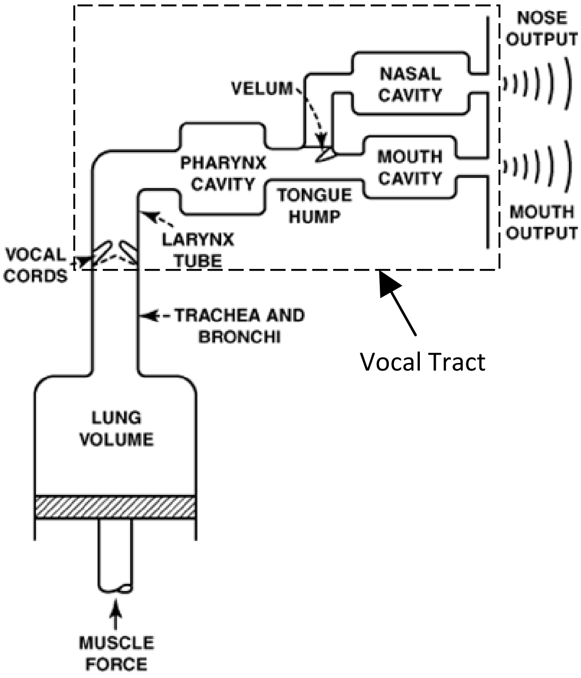

Fig. 1: The model of the vocal tract [7].

As VTLN method is used to eliminate inter-speaker variability, this paper focus on the warp factor and warping frequency cutoff implementation effect on the phoneme recognition performance. Section 2 contains the experimental setup for running the experiment. The recognition results are presented in Section 3, following by the results elaboration in Section 4 before conclusion in Section 5.

2. Experimental Setup

2.1. Preparation

The experiment begins with the speech corpus and toolkit preparation for phoneme recognition. Phoneme recognition approach is considered to be a very delicate recognition task which focuses on recognizing phonemes from every speech corpus files. This approach enables the observation task on the actual recognition performance level of every phoneme in a sentence.

TIMIT Acoustic-Phonetic Continuous Speech Corpus contains a total of 6300 sentences, with 10 sentences spoken by each of 630 speakers with different sex from 8 major dialect regions of the United States. The dialect sentences (SA sentences) are more dialectal variants compare to other sentences [1,9,10] in TIMIT, and thus are removed from the experiment setup to ensure the experiment is free from dialectal variants. After the exclusion of dialectal variants, a total remaining data of 5040 sentences are divided into training data and testing data. Total training data to be used to conduct the experiment are 3696 sentences, while total testing data are 1344 sentences. TIMIT transcriptions are included together with the speech data, and consist of 61 phonemes. Due to TIMIT speech corpus in waveform, it is necessary to convert the speech corpus into digital form. Mel-Frequency Cepstral Coefficient (MFCC) is widely use audio representation form as it gives good discrimination between speech variations [5] and takes human perception sensitivity with respect to frequencies into consideration [14], making MFCC less sensitive to pitch. MFCC also offers better suppression of insignificant spectral variation in higher frequency band and able to preserve sufficient information for speech and phoneme recognition with a small number of required coefficients [13].

The conversion from waveform to MFCC (Figure 2) is done by using HTK, a command prompt based toolkit selected as a medium to train and test out every TIMIT speech corpus to obtain the highest possible recognition performance from every different setting. MFCC conversion also enables the HTK to read and process the input properly [5]. Each HTK setting depends on its configuration setup for transforming, training and testing the speech. The implementation of VTLN can be done from Mel Filterbank Analysis (Figure 3) through this HTK configuration setup.

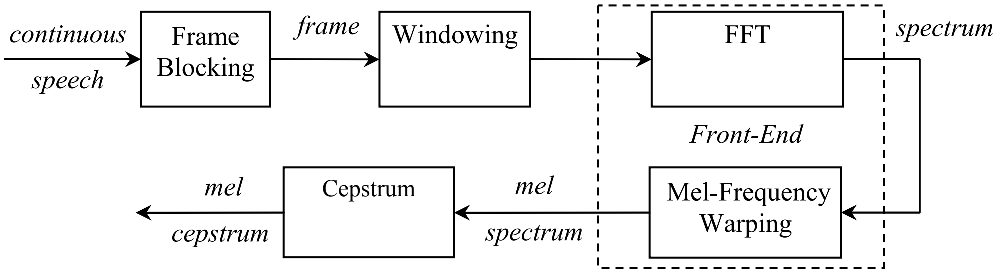

Fig. 2: MFCC conversion process flow.

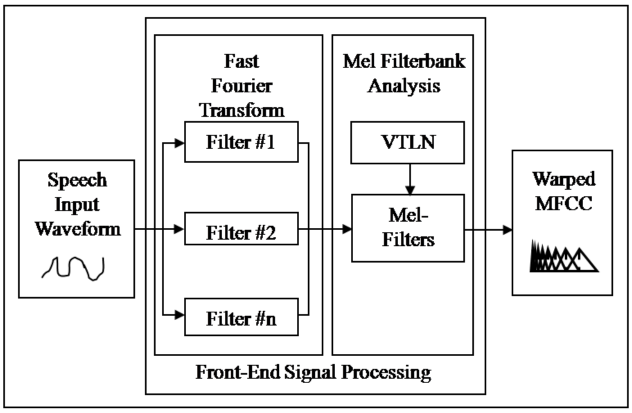

Fig. 3: Conversion from waveform to MFCC with VTLN (Derived from [3], [4] and [6]).

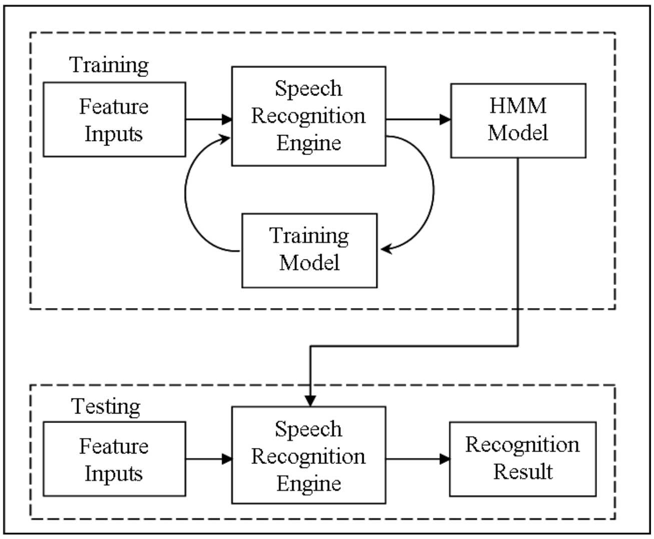

Fig. 4: Experimental training and testing flow (Derived from Young et al., 2006 [5]).

2.2. Training and Testing

An experimental flow diagram is drawn to briefly present the way this experiment is conducted, as showed in Figure 4. This experimental flow is to be repeated done for each VTLN implementation.

VTLN normalizes the speech signal and attempts to reduce the inter-speaker variability in speech signal by compensating for vocal tract length variation among speakers [8]. In HTK configuration setup, VTLN setting consists of warp factor parameter, lower and upper warping frequency cutoff parameter. These three parameters control the minimum and maximum frequency range subjected to be warped at factor α. The main approach of this VTLN implementation is to rescale the frequency axis within the defined frequency boundary on the speech spectrum according to the specified warp factor, α. This type of readjustment, also called piecewise linear warping, can be either stretching or compressing the speech spectrum at warp factor α. Since the suitable warp factor is unknown in this experiment, a range factor of 0.5 to 2.0 is used in trial-and-error approach, with the increment of 0.02 for each experiment. This warp factor range limit is reasonable as the spectrum will lost its information either after being compressed by half with factor 0.5 or after being stretched twice the size with factor 2.0. The trial-and-error approach also requires high computational resource in order to obtain every recognition performance results. By properly run experiment for each VTLN setting applied in TIMIT speech corpus, this will help to evaluate the result and identify the best setting for TIMIT speech corpus.

3. Results

The experiment is focused on phoneme recognition, so the recognition performance result is measured by Phoneme Error Rate (PER). Both language model, Unigram and Bigram, are used in this experiment and the recognition performance are recorded as well for comparison.

The best recognition performance result is selected among every single experiment from selected language model, as shown in Table 1 and 2. PER can be calculated from the number of substitution errors (S), deletion errors (D), insertion errors (I), total phoneme (N) and total correct (H). Each value from the tables below is calculated based on these equations below:

$$H = N - S - D\,. \tag{1}$$

$$Corr = \frac{H}{N} \times 100\%\ . \tag{2}$$

$$Acc = \frac{H - I}{N} \times 100\%\ . \tag{3}$$

$$PER = 100\% - Acc\ . \tag{4}$$

Table 3 and 4 summarize the lowest PER achievement from 2 different language models, after VTLN implementation. Starting with warp factor 1.0 representing non-VTLN implementation, the initial recognition performance is 38.83% for Unigram language model, and 29.57% for Bigram language model.

Table 1: Phoneme Recognition Result with Warp Factor 1.38 for Unigram Language Model

N	H	D	S	I	Accuracy Rate	Correct Rate
51681	38473	2393	10815	6477	61.91%	74.44%

Table 2: Phoneme Recognition Results with Warp Factor 1.40 for Bigram Language Model

N	H	D	S	I	Accuracy Rate	Correct Rate
51681	38230	4580	8871	1431	71.20%	73.97%

Table 3: Recognition Performance for Unigram Language Model

Warp Factor	Upper Warp Cutoff Frequency (Hz)	Accuracy Rate	Phoneme Error Rate
1.00	-	61.17%	38.83%
1.38	3800	61.91%	38.09%

Table 4: Recognition Performance for Bigram Language Model

Warp Factor	Upper Warp Cutoff Frequency (Hz)	Accuracy Rate	Phoneme Error Rate
1.00	-	70.43%	29.57%
1.40	5000	71.20%	28.80%

4. Discussion

Warp factor of 1.00 equates to non-VTLN implementation, making this warp factor value suitable to be treated as controlled variable. The performance result for warping factor 1.00 is used as initial reference to observe the performance changes for different warping factors. During the time this experiment is conducted, lower warp frequency cutoff value is fixed at 300Hz as it won’t affect much to recognition performance.

Bigram language model shows better phoneme recognition performance compare to Unigram language model, with more than 24% performance improvement. It is due to better state matching reference which compares the trained HMM model with two states of a phoneme test data instead of one state.

Another noticeable similarity between two language models is the better accuracy rate with warp factor above 1.00. As the experiment setup is done with speaker independent mode in mind, the accuracy rate achieved is considered as the averaged recognition performance regardless of speaker’s gender.

5. Conclusion

This experiment shows that phoneme recognition performed well on TIMIT speech corpus when the warp factor value is more than 1.00. HTK performed best in Bigram language model when the warp factor is 1.40, with 28.8% PER. Although trial-and-error approach gives precise identification on the best warp factor, further experiment need to be done on word level recognition performance, by applying the same setting for phoneme recognition.

6. References

[1] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, N.L. Dahlgren, *DARPA TIMIT Acoustic-phonetic Continuous Speech Corpus*, U.S. Department of Commerce, NIST, Gaithersburg, MD, 1993.

[2] Rabiner, L., Juang, B-H. *Fundamentals of Speech Recognition*, Prentice-Hall International, 1993.

[3] Lee, L., Rose, R.C. *Speaker normalization using efficient frequency warping procedures*, Proc. IEEE ICASSP 96. 1, 353-356. 1996.

[4] Lee, L., Rose, R.C. *A Frequency Warping Approach to Speaker Normalization*, IEEE transactions on speech and audio processing. 6(1), 49-60. 1998.

[5] Young, S. et al. *The HTK Book*, Cambridge University Engineering Department. (8th ed.). 2006

[6] Zhan, P., Waibel, A. *Vocal Tract Length Normalization for Large Vocabulary Continuous Speech Recognition*, CMU-CS-97-148, Carnegie Mellon University, Pittsburgh, PA. May. 1997

[7] Flanagan, J.L. *Speech Analysis and Perception*. (2nd ed.) Verlag, Berlin: Springer. 1965.

[8] Giuliani, D., Gerosa, M. and Brugnara, F. *Improved Automatic Speech Recognition through Speaker Normalization*. Computer Speech & Language, 20 (1), pp. 107-123, Jan. 2006.

[9] Lee, K.F., Hon, H.W. *Speaker-independent phone recognition using hidden Markov models*. IEEE Trans. Acoustics, Speech and Signal Processing 37(11), pp. 1641-1648, 1989.

[10] Müller, F., Mertins, A., *Robust speech recognition based on a certain class of translation-invariant transformations*, in LNCS, Vol 5933, pp. 111-119, 2010.

[11] Müller, F., Mertins, A., *Invariant Integration Features Combined with Speaker-Adaptation Methods*, in Proceedings of Int. Conf. Interspeech 2010, 2010.

[12] Liu, M., Zhou, X., Hasegawa-Johnson, M., Huang, T.S., Zhang, Z.Y., *Frequency Domain Correspondence for Speaker Normalization*, in Proc. INTERSPEECH, 2007, pp. 274-277.

[13] Davis, S.B., Mermelstein, P., *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences*, in IEEE Trans. on Acoustics, Speech and Signal Processing, Vol 28, pp. 357-366, 1980.

[14] Roger Jang, J.S., *Audio Signal Processing and Recognition*, available at the links for on-line courses at the author's homepage at http://www.cs.nthu.edu.tw/~jang.