An Approach to Vocal Tract Length Normalization by Robust Formant

A. Kabir, J. Barker, and M. Giurgiu

***Abstract*—**Spectrum pattern of the same phoneme could be quite different for individual speakers due to physical and linguistic difference. Without applying appropriate computational technique on the frequency axis, the inter-speaker variation will reduce the modeling efficiency and result in poor recognition performance. In this paper, a formant-driven framework is proposed which is based on by modifying formant pattern model in order to compute normalization factor of a given speaker. Experiments on GRID corpus clearly show the effectiveness of this method.

***Keywords*—**Vocal Tract Length Normalization, Frequency Warping, Formant Pattern Model.

I. Introduction

Current state of art automatic speech recognition (ASR) systems has one of the major challenges due to the inter-speaker variation, context, and environments [1, 2]. But only considering inter-speaker variations, physiology difference (vocal tract shape and length, etc.) and linguistic difference (accent and dialect, etc.) could be the main reasons to it. It is also very clear that performances of speaker dependent (SD) ASR systems are better than speaker independent (SI) ASR systems because of these variations. Normally SD ASR systems have the half of error rates in compare to SI ASR systems for the same task and it is particularly important in SI ASR systems which are designed for dealing with any arbitrary unknown speaker in applications such as directory assistance [5]. It is reported in [2] that the spectrum pattern for the same phoneme of two speakers can be very different due to physiology difference and linguistic difference. Therefore, fair enough to assume that performance of SI ASR systems could be substantially improved if inter-speaker variations are minimized by applying appropriate computational techniques.

It is well known that speech signal carries information about vocal tract length (VTL): for example, formant frequencies of vowels decrease as the VTL increases. Vocal Tract Length Normalization (VTLN) is very common in ASR systems in order to minimize inter-speaker variations. VTLN is a technique which scales the frequency axis of the acoustic feature vectors by introducing a warping or normalization factor so that the observations are alike across all speakers [2, 8]. VTLN is able to reduce the inter-speaker variation after warping the spectrum and this is especially valuable in gender independent systems because on average VTL is 2-3cm shorter for female speakers than male speaker, resulting formant frequencies of female speakers are 15-20% higher than male speakers [2, 3, 8].

There are many algorithms available in the literature to estimate warp factor or normalization factor in order to reduce the inter-speaker variation. These methods can be categorized into two classes: model based normalization and feature based normalization. The most common method for finding warp factors for model based normalization operates on the maximum likelihood (ML) criterion to choose a warp factor that gives a speaker's warped observation vectors the highest probability [2, 8]. The likelihoods can be computed using the recognizer's phone models. On the other hand, feature based normalization predicts warp factors by observing more direct parameters of speech acoustics, such as formants (resonant frequencies of the vocal tract). The first and second formants can be modeled by vowel-specific distributions [2, 8].

There are mainly three different warping functions are used: linear warping, nonlinear warping and piecewise linear warping. In linear warping, one parameter will determine the global warping which may not be sufficient to compensate the total variation of different speakers [2]. Nonlinear warping and piecewise linear warping are proposed to further improve the efficiency of the recognizer. But all of these normalization methods are essentially maximizing the likelihood of utterance given a model [2].

In this paper, we are proposing a formant based method in order to apply VTLN technique into a large speech corpus known as GRID corpus. Normalization factor is computed with the robust reliable formant estimation method corresponding to a given speaker, the inter-speaker variation can be reduced prior to acoustic modeling procedure which will increase the modeling efficiency. Experimental results on GRID corpus shows the effectiveness of this method.

The rest of the paper is organized as follows: paper is organized as follows: First we describe the experimental corpus, then we present the method used to compute

Manuscript received August 12, 2010. This work was supported by the the EC-funded Project Marie Curie Research Training Network MRTN-CT-2006-035561 S2S ("Sound to Sense", www.sound2sense.eu).

A. Kabir, is with the Department of Telecommunications, Technical University of Cluj-Napoca, 400027 Cluj-Napoca, Romania. (phone: +40-264-401807; e-mail: ahsanul.kabir@com.utcluj.ro).

J. Barker, is with the Department of Computer Science, Sheffield S1 4DP, United Kingdom. (e-mail: j.barker@dcs.shef.ac.uk).

M. Giurgiu, is with the Department of Telecommunications, Technical University of Cluj-Napoca, 400027 Cluj-Napoca, Romania. (e-mail: mircea.giurgiu@com.utcluj.ro).

normalization factor as well as to perform VTLN by speaker adaptive training. Subsequently we present and interpret the experimental results, and the final section draws a conclusion.

II. SPEECH MATERIALS

Speech data has been taken from GRID corpus for this experiment which is a large multi talker audio visual sentence corpus to support joint computational-behavioral studies in speech perception and automatic speech recognition [4]. It contains a total of 34,000 sentences of high quality audio and video (facial) recordings, 1000 sentences spoken by each of 34 speakers (18 male speakers, 16 female speakers). All speak British English as their first language. All but three participants had spent most of their lives in England and together encompassed a range of English accents [4]. Two speakers grew up in Scotland and one was born in Jamaica. Grid provides a greater variety and is large enough to meet the training requirements of ASR systems.

III. Vocal Tract Length Normalization

It can be noted that model based normalization such as ML approach is computationally expensive where we focused on feature based normalization approach which is computationally economic and based on the fixed formant pattern model.

A. *Formant Pattern Model*

The fixed formant pattern model is very simple; a single parameter for each formant and a single scalar value related to the individual speaker's formants to those of the population [9]. Mathematically, the formant pattern model is given by,

$$\lambda_v^i = a_i \lambda_{ref}^v \tag{1}$$

Where, v is the vowel type, i is the individual speaker, λ_v^i is the wavelengths of the formant vector of the individual speaker for vowel v, λ_{ref}^v is the wavelengths of the average formant vector of the entire population for the vowel v, the scalar a_i is the length of individual speaker's vocal tract with respect to the average of the population.

B. *Warping Factor Estimation*

Warping factor is estimated based on the fixed formant pattern model by the following way.

According to the fixed formant pattern model,

$$F_{ref}^v = \alpha_i F_i^v$$

$$\Rightarrow F1_{ref}^v = \alpha_i F1_i^v \tag{2}$$

$$\Rightarrow F2_{ref}^v = \alpha_i F2_i^v \tag{3}$$

Where, F1 and F2 are the first two formants for the reference speaker, and the individual speaker, i of the vowel type, v. And α_i is the normalization factor of the individual speaker, i.

Now the Euclidean distance for the vowel type, v between the vowel points of the reference speaker and the individual speaker after applying VTLN is given by,

$$d = \sqrt{(F1_{ref} - \alpha_i F1_i)^2 + (F2_{ref} - \alpha_i F2_i)^2}$$

$$\Rightarrow d^2 = a\alpha_i^2 - 2b\alpha_i + c \tag{4}$$

Where,

$$a = (F1_i)^2 + (F2_i)^2 \tag{5}$$

$$b = F1_i F1_{ref} + F2_{ref} F2_i \tag{6}$$

$$c = (F1_{ref})^2 + (F2_{ref})^2 \tag{7}$$

Normalization factor α_i is given at the minimum distance between the vowel points. Therefore differentiating (5) with respect to α_i yields,

$$2d * d' = 2(a\alpha_i - b) \tag{8}$$

$$\Rightarrow \alpha_i = {}^{b}\!/_{a} \tag{9}$$

Substituting, $d' = 0$ into (6).

Equation (7) gives the normalization factor for a specific vowel type, v of a specific speaker, i. Then for each vowel of the specific speaker, normalization factor is computed. Ideally, all normalization factors should be same but in practice, deviation exists mainly due to the difficulties of formant estimation. That is why mean of these normalization factors is taken as the normalization factor of the specific speaker.

C. *Speaker Adaptive Training & Testing*

Speaker Adaptive Training (SAT) is a technique used to train SI acoustic models that integrates normalization factor. We applied VTLN in both during training and testing phase which have been performed by the following steps,

1. Firstly, training and testing have has been carried out by the non-normalized features. We labeled this procedure as the baseline system.
2. Then normalization factor has been computed through by the reliable formant frequencies and training and subsequently, testing has been carried out, and labeled as VTLN by reliable formants.
3. Finally, normalization factor has been computed through the parametric approach by the highly reliable formant frequencies and subsequently, training and testing has been carried out and labeled as VTLN by highly reliable formants.

IV. Experimental Results

A. *Formant Measurements*

Formant frequencies have been estimated for the entire GRID corpus by three widely used fully automatic formant tracking algorithms (classical LPC method, burg algorithm implemented in popular and widely used software PRAAT, and Auto Regressive method). Then reliable formants (RF) and highly reliable formants (HRF) are determined by comparison with formant frequencies provided by Detering [6] and Hawkins [7]. Robust reliable and highly reliable formant estimation technique is based on a data refinement algorithm where the algorithm chose the most likely formants estimated by these three widely used algorithms. Figure 1 shows the vowel space of highly reliable formants and ellipse has been drawn by their standard deviation around the average formant frequencies of the entire population.

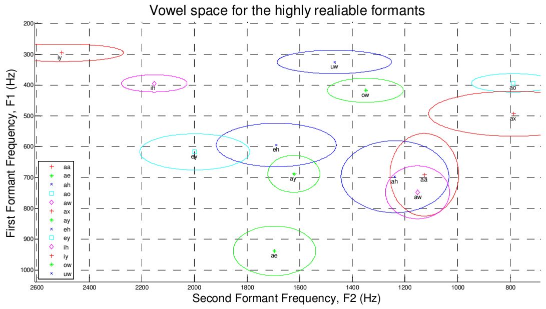

Fig 1. vowel space for highly reliable formants in comparison to Hawkins [7]

B. *Warping Factor Estimation*

A VTL warping factor for each speaker has been estimated by applying equation (7) using reliable formants and highly reliable formants of all vowels. For each vowel independently, average formant frequencies for F1, and F2 are computed for the entire population in order to provide the reference point for formant frequency warping. A normalization factor is applied to each formant which provides the minimum distance between formant frequencies of an individual speaker and the reference speaker. The average of the normalization factor calculated over all vowels is taken as the normalization factor of the specific speaker.

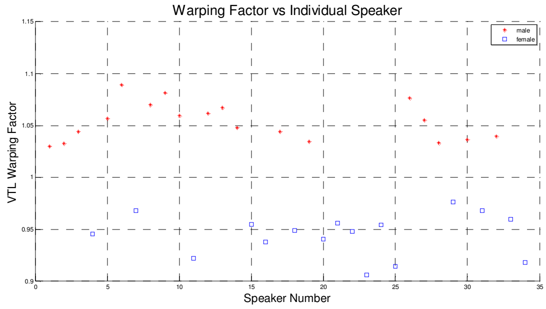

Fig 2. warping factor for each speaker in the GRID corpus estimated from the highly reliable formants which are obtained in comparison to Hawkins [7]

Figure 2 shows normalization factor of each speaker of the GRID corpus plotted against speaker number. Separation between the male speakers and female speakers could be easily identified. It could be noted that reference speaker has a warping factor of one.

C. *Vowel Normalization*

Figure 3 shows the normalized vowel space after applying VTLN to each speaker of the GRID corpus. Vocal tract length geometry is different between adult male speakers and adult female speakers. Therefore, it results a scatter distribution of formant frequencies when both type of speakers are present in the data set. But if vocal tract shapes are normalized to a reference speaker, a compact distribution of formant frequencies would also result. It is clear from Figure 5 that the formant frequency distribution is compact.

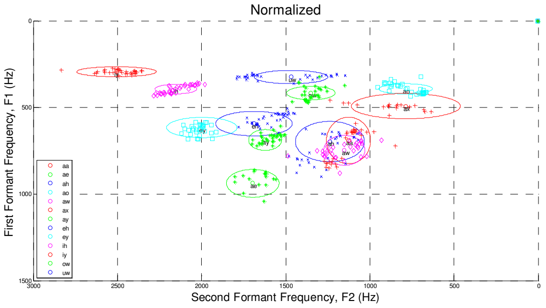

Fig 3. normalized vowel after applying VTLN

D. *Speech Recognition Experiment*

Figure 4 shows the performance of the baseline system (without normalization), applying VTLN where warping factor is estimated by reliable formant frequencies determined by in comparison to Deterding [6] and Hawkins [7], and applying VTLN where warping factor is estimated by highly reliable formant frequencies determined by in comparison to Deterding [6] and Hawkins [7]. This figure only shows the recognition rate of all the phonemes when testing is carried out in a speaker independent manner.

TABLE I. Recogntion Rate of Baseline System, and After Applying VTLN for SI ASR and GD ASR

		Baseline	VTLN			
			By Reliable Formants		By Highly Reliable Formants	
			RF [7]	RF [8]	HRF [7]	HRF [8]
	SI	77.94	79.30	79.33	79.21	79.52
GD	Male	79.49	80.02	79.93	80.04	79.97
	Female	82.97	83.18	83.12	83.46	83.32

For Speaker Independent (SI) ASR, Baseline system has a recognition rate of 77.94%. Recognition rate is improved by 1.4% when VTLN is applied and normalization factor is estimated by reliable formant frequencies. On the other hand, when normalization factor is estimated by highly reliable formant frequencies recognition rate is improved by 1.7% in the best case. GRID is a very large speech corpus and an improvement of 1.7% is not negligible. Moreover, a substantial difference in recognition rate does not exist in terms of computing normalization factor by reliable formant frequencies and highly reliable formant frequencies.

For Gender Dependent (GD) ASR, Baseline system has a recognition rate of 79.49% for male speakers and 82.97% for female speakers. Recognition rate is improved about 0.50% for both kinds of speakers after applying VTLN. Though GD ASR outperformed SI ASR in each case but it is fair enough to say that there is no noticeable improvement in recognition rate in GD ASR systems after applying VTLN. It is because there would be little variation in the VTL pattern among male speakers as well as among female speakers.

V. Conclusion

In this paper, we presented a feature based normalization approach to robust VTLN in order to reduce inter-speaker variation that takes advantage of the first two formant frequencies in order to compute warping factor of each speaker. Reliable formant frequencies and highly reliable formant frequencies are also taken into consideration to compute warping factor. Though VTLN improves the recognition rate in compare to baseline system but we found that there is no substantial difference in recognition rate when normalization factor is computed by reliable formant frequencies and highly reliable formant frequencies. This method could be applied as it reduces the computational requirements.

Acknowledgement

This work has been carried out in the frame of the EC-funded Project Marie Curie Research Training Network MRTN-CT-2006-035561 S2S (“Sound to Sense”, www.sound2sense.eu).

References

[1] P.Zhan, and A.Waibel. “Vocal Tract Length Normalization for Large Vocabulary Continuous Speech Recognition,” Technical Report, Department of Computer Science, Carnegie Mellon University, USA, 1997.

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

[3] G.Garau, S.Renals and T.Hain. “Applying Vocal Tract Length Normalization to Meeting Recordings,” in *Proc. INTERSPEECH*, 2005, pp. 265-268.

[4] M. Cooke, J. Barker, S. Cunningham and X. Shao, “An audio-visual corpus for speech perception and automatic speech recognition.” *Journal of the Acoustical Society of America*, vol. 120, 2006.

[5] S.Umesh, D.R.Sanand and G.Praveen, “Speaker-Invariant Features for Automatic Speech Recongnition.” in *Proc. of the International Joint Conference on Artificial Intelligence*, 2007, pp. 1738-1743.

[6] D.Deterding. “The Formants of Monophthong Vowels in Standard Southern British English Pronunciation,” Journal of the International Phonetic Association, vol. 27, pp. 47–55, 1997.

[7] S.Hawkins and J.Midgley. “Formant frequencies of RP monophthongs in four age groups of speakers,” Journal of the International Phonetic Association, vol. 35, pp. 183–199, 2005.

[8] A.Faria and D.Gelbart. “Efficient Pitch-based Estimation of VTLN Warp Factors,” in *Proc. INTERSPEECH*, 2005, pp. 213-216.

[9] R.E.Turner, T.C.Walters, J.J.Monaghan and R.DPatterson, “A statistical, formant-pattern model for segregating vowel type and vocal-tract length in development formant data.” *Journal of the Acoustical Society of America*, vol. 125, 2009.

[10] J. Barker and M. Cooke, “Modeling Speaker Intelligibilty in Noise.” *Speech Communications*, vol. 49, 2007, pp. 402-417.

[11] H.Wakita. “Normalization of Vowels by Vocal-Tract Length and Its Application to Vowel Identification,” *IEEE Transaction on Acoustics, Speech and Signal Processing*, vol. ASSP-25, 1977, pp. 183-192.