BAYESIAN CONSTRAINED FREQUENCY WARPING HMMS FOR SPEAKER NORMALISATION

[1, 2] *Ching-Hsiang Ho* [1] *Saeed Vaseghi* [1] *Aimin Chen*

[1] *The Queen's University of Belfast, Northern Ireland*
[2] *Fortune Junior College of Technology and Commerce, Taiwan, R.O.C.*

ABSTRACT

This paper presents a Bayesian constrained frequency warping technique. The Bayesian approach provides for inclusion of the prior information of the frequency warping parameter and for adjusting the search range in order to obtain the best warping factor dependent on HMMs. We introduce novel frequency warping (FWP) HMMs which are different warped versions of HMMs. Instead of frequency warping of the input speech we warp the spectrum of the HMMs. This is equivalent to HMMs which have both time and frequency warping capabilities. Experimentally FWP HMMs outperform the conventional constrained frequency warping approach. Furthermore, the best warping factor is estimated in two stages, a coarse stage followed by a fine stage. This method efficiently gauges the optimal warping factor and normalises the FWP HMMs.

1. INTRODUCTION

A major source of inter-speaker variability in hidden Markov model (HMM) based recognition is due to the variations of shape and length of the vocal tract among different speakers. Vocal tract variability results in a broadening of speaker independent HMM probability models and a mismatch between the distributions of the training utterances and test data. The effect of vocal tract length variation is a shift in the spectrum of the speech. In [2] two simple vocal tract models are considered, the uniform tube and the Helmholtz resonator. In the uniform tube the formant frequencies of utterances for a given sound are inversely proportional to the length of the vocal tract. The scaling of the formant frequency in the uniform model is consistent with linear frequency warping. However, in the Helmholtz resonator model, a good approximation for the closed front vowel, the formant frequencies are inversely proportional to the square root of the vocal tract length. Furthermore, the higher frequency regions for these vowels show more sensitivity to the variation of the vocal tract length. Therefore, the scaling of the frequency axis imposed by a change in vocal tract length is dependent on the configuration of the vocal tract (the phoneme) and also the frequency regions.

Recently, several maximum likelihood based frequency warping procedures have been proposed to reduce the speaker dependent variability via front-end signal processing [1][2][3][4]. These frequency warping methods linearly or nonlinearly re-scale the frequency axis to reduce the variations between formant frequencies. A maximum likelihood estimation is used to find the optimal warping parameters which maximise the likelihood between a set of reference statistical models and the input utterances. Since the warping transformation is employed in the front-end stage, it is hard to find a closed-form solution for the ML criteria. Therefore, a grid search is used to exhaustively search the optimal warping parameters over an extended space of utterances.

In this paper, a Bayesian constrained frequency warping method is presented where the prior information of the warping factor is incorporated to efficiently search for the optimal factor. Firstly, a by-product of the iterative normalisation procedure in training is a set of probability models for the distribution of the frequency warping parameters for each HMM. These probability models are then employed for a Bayesian speaker normalisation of the training and the test data.

In a second approach to Bayesian frequency warping, instead of warping the input speech, frequency warping (FWP) HMMs are employed to model the vocal tract variations. FWP HMMs are a set of extended HMMs where each HMM is associated with a range of warping parameters and are estimated by maximising the likelihood of the extended observations. The best warping factors for the input speech can be obtained by searching FWP HMMs over all utterances. It is equivalent to HMMs which have both time and frequency warping capabilities.

Furthermore, the two methods above are combined in a two-step iterative procedure to implement the Bayesian constrained frequency warping. In this procedure the most likely range for the warping parameters is firstly estimated by searching FWP HMMs. Then, the optimal warping factor is estimated within the optimised constrained range. Therefore, both the efficiency of the HMM search and the precise warping of the observations are encapsulated in the novel approach.

In Section 2 the maximum likelihood based frequency warping is described. Section 3 proposes Bayesian constrained FWP HMMs. In Section 4 a novel efficient frequency warping technique is presented. Section 5 presents some experimental results. Section 6 concludes this paper.

2. ML BASED FREQUENCY WARPING

The maximum likelihood based frequency warping [2][3][4] is a vocal tract normalisation approach using HMM based speech recognition. The advantage is that it is easy to incorporate this frequency warping method into an automatic speech recognition system.

During both training and recognition, the optimal linear frequency warping factor is estimated by maximising the likelihood of the utterances with respect to a set of given HMMs. Suppose that O_i denotes a set of utterances spoken by speaker i, S_i denotes the corresponding state sequence transcriptions for the utterances, and $[\lambda]$ denotes a set of HMMs. The optimal warping factor, $\hat{\alpha}$, is estimated from a set of N discrete values within a constrained range and defined as

$$\hat{\alpha} = \arg\max_{\alpha} P\left(O_i \middle| \alpha, \lambda, S_i\right) \quad (1)$$

Since finding a closed-form solution for Equation 1 is a nontrivial exercise, a grid search procedure is used. The procedure is shown in Figure 1 and is described as follows:

1. For an utterance O_i, given a set of HMMs, $[\lambda]$, the ML state sequence transcription, S_i, is obtained using the Viterbi search.
2. N sets of warped utterances are obtained by warping the utterance, O_i, using a set of N discrete warping factors, α_1 *to* α_N,.
3. Each set of warped utterances from Step 2 is aligned with the corresponding state sequence transcription, S_i, from Step 1 while the joint probability of all frame vectors is obtained from the pdfs of the mixture states.
4. The best warping factor is the one which maximises the likelihood of the corresponding set of warped utterances.

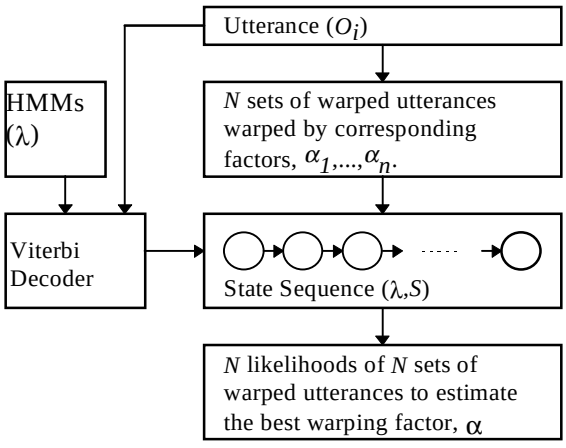

Figure 1: The grid search procedure

The goal of the training procedure is to reduce the inter-speaker variability by warping the frequency scale of the utterances. A set of narrower distributions is obtained by retraining the normalised utterances. During recognition, the optimal warping factor estimated by the same MLE procedure is used to remove the mismatch between the normalised HMMs and the test utterances. The training and testing procedures are described as follows.

Training procedure:

1. The MLE is used to estimate the best warping parameters.
2. All training utterances, warped by the optimal warping factor, are then used as the retraining database.
3. The retrained models are called normalised models.
4. This procedure is iteratively executed until a certain convergence condition is achieved.

Testing procedure:

1. The MLE is used to estimate the best warping parameters.
2. All test utterances warped by the corresponding optimal warping factors are then recognised against the normalised models.

3. BAYESIAN CONSTRAINED FWP

The development of the Bayesian constrained frequency warping method can be described in three stages. Firstly, the prior information of the warping factor is investigated. Secondly, a more representative statistical model and efficient search of the warping factor is presented. Thirdly, we combine the two methods mentioned above to obtain a more optimal solution.

3.1. Constrained Bayesian Methods

In this section we propose several constraints and prior information of the warping factor for developing constrained Bayesian approaches. The conventional ML frequency warping method iteratively uses the exhaustive search to estimate the optimal warping factor. However, some useful information about the warping factor is ignored. In order to make the search more optimal, we try to incorporate the prior pdf of the warping factor for each HMM. Thus, the maximum a posteriori estimation for the warping factor is obtained by solving

$$\hat{\alpha} = \arg\max_{\alpha} P\left(o|\alpha, \hat{\lambda}\right) P\left(\alpha|\hat{\lambda}\right) \quad (2)$$

where the distribution of the warping parameter is incorporated into Equation 1 as a priori information.

The simplest case is the ML based frequency warping. In the grid search the vocal tract variation which is within 25% is used as the prior knowledge to constrain the warping factor. In this case the prior is an uniform pdf within the grid search range. However, in a gender-unbalanced training database, such as TIMIT, the mean of the warping factors is biased so that it is inappropriate to use a fixed search range. By incorporating prior knowledge of the warping factor, the search range becomes adaptive. Besides, since the distribution of the warping factor is not uniform, the step size can be made adjustable using the prior probability distribution.

Based on the parametric model of the vocal tract [1], the warping factor is a linearly scaling factor between the change of the vocal tract length/shape and the variation of the formant frequencies. Under this condition the warping factor becomes phoneme dependent. To take advantage of the characteristics of the distribution of the warping factor for each phoneme, the distribution is incorporated as the prior pdf. Although there is a certain amount of information associated with the distribution of the warping factor and an improvement is attained, a more informative model of the warping factor is still needed.

3.2. Frequency Warping HMMs

FWP HMMs are proposed and incorporated to model the warping factors. FWP HMMs are a set of extended HMMs, $[\hat{\lambda}^{\alpha'}]$. Given a set of reference HMMs, $[\lambda^{\alpha'}]$, FWP HMMs are estimated by maximising the likelihood of the extended observation sequence, o^{A}, as

$$\hat{\lambda}^{\alpha'} = \arg\max_{\lambda^{\alpha'}} P\left(o^{A} \middle| \lambda^{\alpha'}\right) \tag{3}$$

where A denotes the whole discrete set of warping factors. Thus, the optimal warping factors for input utterances can be estimated by

$$\hat{\alpha}' = \arg\max_{\alpha'} P\left(o \middle| \hat{\lambda}^{\alpha'}\right) \tag{4}$$

where the Forward-Backward algorithm or Viterbi decoder can be used.

A training procedure for Equation 3 is developed in which FWP HMMs are obtained by using all training utterances and their extended versions. The extended utterances are generated by warping the training utterances using the set of warping factors. The procedure is described as follows:

1. To generate the extended training utterances, the frequency scale of all training utterances is warped by a discrete range of warping factor.
2. The state sequences of the unwarped utterances are obtained by forced-alignment where a set of HMMs is given. The set of HMMs can be SI HMMs at the beginning or FWP HMMs during the training iterations.
3. The segmental labels obtained from Step 2 are applied to all the corresponding warped utterances.
4. The optimal warping factor for each phone segment is estimated by maximising the likelihood of the observation sequence, given SI HMMs or FWP HMMs.
5. Each phone in the extended training utterance is re-labeled by a new name which is the original name attached with the estimated warping factor of the segment from Step 4.
6. Using the Baum-Welch algorithm, FWP HMMs are estimated by maximising the extended training utterances. For each phone model of the initial SI HMMs a set of warped models are obtained.
7. The further training iteration can be used to optimise the FWP HMMs.

3.3. A Novel Warping Factor Estimation

According to the experimental results of the ML based frequency warping, the supervised frequency warping is better than the unsupervised case particularly when warping in the subword segment. In supervised normalisation the phone transcriptions are employed during the estimation of the state sequence. However, as we warp utterances frame by frame in the supervised case, the recognition rate increases slightly but the accuracy is degraded. It seems that an improvement can be achieved if we use soft segmentation instead of hard segmentation in frequency warping training. Therefore, before searching for the optimal warping factor, we incorporate the FWP HMMs search to estimate a coarse factor.

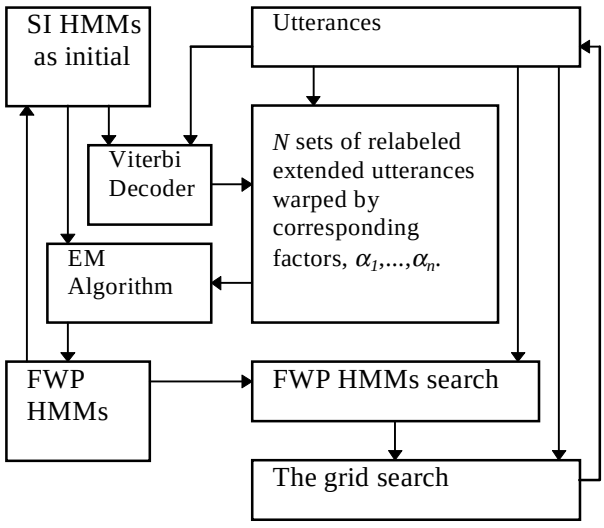

Figure 2: The warping factor estimation algorithm

The iterative procedure for the novel warping factor estimation is shown in Figure 2 and described as follows. We define a

coarse warping factor, α_r, and a fine warping factor, α_f. The optimal warping factor, $\hat{\alpha}$, is obtained by adding the two factors together as in Equation 5. The optimal coarse warping factor, $\hat{\alpha}_r$, defined as Equation 6, is estimated by maximising the likelihood of the observation sequence, *X*, given FWP HMMs, λ^{α_r}. The optimal fine warping factor, $\hat{\alpha}_f$, is estimated by maximising the likelihood of the observation sequence, X^{α_f}, given estimated FWP HMMs, $\lambda^{\hat{\alpha}_r}$, as Equation 7.

$$\hat{\alpha} = \hat{\alpha}_r + \hat{\alpha}_f \tag{5}$$

$$\hat{\alpha}_r = \arg\max_{\alpha_r} P\left(X \middle| \lambda^{\alpha_r}, \alpha_r\right) \tag{6}$$

$$\hat{\alpha}_f = \arg\max_{\alpha_f} P\left(X^{\alpha_f} \middle| \lambda^{\hat{\alpha}_r}, \alpha_f\right) \tag{7}$$

During training, the optimal warping factor is estimated using the new warping factor estimation procedure. During testing, the normalised FWP HMMs can be employed to recognise the test utterances without predetermining the warping factors for the utterances.

4. EXPERIMENTS

To study various characteristics of the warping factor, we investigate the frequency warping of different segments. Generally speaking, when the warping factor varies across a smaller segment, the normalisation contributes a better performance.

Experiments were based on the TIMIT speech database using monophone HMMs. The speech features consisted of 13 MFCCs supplemented with the 1st and 2nd differentials. The constraint of linear frequency warping factor is from 0.88 to 1.12 with steps of 0.02. With a single Gaussian per state the baseline recognition rate is 58.10%. Table 1 and 2 compare recognition results with respect to different warping segments. These are the observations from each speaker (SPWF), each sentence (SEWF), each phone segment (PHWF) and each state segment (STWF). Firstly, the supervised (S) normalisation results in a better performance than the unsupervised (U) normalisation. Secondly, when the warping factor varies across the sub-word segment, the improvement is significant. It shows that frequency warping factor is speaker dependent as well as phoneme dependent. Thirdly, since using supervised normalisation for SPWF does not improve recognition rate, it is likely that the model alignment is less important when the warping factor is speaker dependent only.

Table 1: Normalisation in recognition (1 Gaussian per state)

	SPWF	SEWF	PHWF
S	58.24%	58.92%	60.95%
U	58.20%	58.73%	58.28%

Table 2: Normalisation in 1-iteration training and recognition with 1 Gaussian per state

	SPWF	SEWF	PHWF	STWF
S	58.53%	59.77%	61.97%	62.52%
U	58.49%	58.56%	59.48%	58.90%

In Table 3 the 8 Gaussian multi-mixture HMMs contribute significant improvement when the normalisation procedure is incorporated into recognition (R) and also the HMM training.

Table 3: Normalisation with 8 Gaussian per state and PHWFs

	BASE	R	T+R
S	69.33%	69.41%	71.14%

Table 4 shows that the HMMs obtained from supervised training are inappropriate to the unsupervised recognition.

Table 4: S/U Training + S/U Recognition

	U+U	S+S	S+U
PHWF	59.48%	61.97%	58.79%

When we apply the FWP HMMs to the normalisation, the recognition rate is 60.97%. It shows that, compared to the S+U normalisation using ML frequency warping, FWP HMM search contributes a significant performance.

5. CONCLUSION

The Bayesian constrained frequency warping approach contributes significant improvements both in efficiency and recognition. FWP HMMs have been successfully employed in the estimation of the frequency warping factor and have improved S+U normalisation by 2%. Also during recognition the search for the phoneme sequence and warping factor can be done simultaneously without warping the input speech. In addition, the novel warping factor estimation procedure will be used to efficiently and precisely to estimate the optimal warping parameters and the normalised FWP HMMs.

6. REFERENCES

1. Ellen Eide and Herbert Gish, (1996) A Parametric Approach to Vocal Tract Length Normalization. *Proc. ICASSP'96*, pp. 346-348.
2. Li Lee and Richard Rose, (1998) A Frequency Warping Approach to Speaker Normalization. *IEEE Trans. Speech, Audio Processing*, Vol6, No. 1, pp. 49-60.
3. Steven Wegmann, Don McAllaster, Jeremy Orloff and Barbara Peskin, (1996). Speaker Normalization on Conversational Telephone Speech. *Proc. ICASSP'96*, pp. 339-341.
4. P. Zhan and M. Westphal, (1997) Speaker Normalization Based on Frequency Warping. *Proc. ICASSP'97*, pp. 1039-1042.