Internal Use Only (非公開)

002

TR-IT-0248

Practical Vocal Tract Length Normalization for Automatic Speech Recognition

マイケル バッキアーニ
Michiel Bacchiani

1997.12

In this report, the background and implementation of speaker normalization using both a likelihood based vocal tract length estimation procedure as well as a formant based method are described. The implementation was done using the ASSM software package described in detail in TR-IT-0147.

©ＡＴＲ音声翻訳通信研究所

©ATR Interpreting Telecommunications Research Laboratories

Contents

1 Introduction 2

2 Speaker Normalization Algorithms 3
- 2.1 Frequency warping 3
- 2.2 Formant based warp estimation 5
- 2.3 Maximum Likelihood Normalization using a Mixture Model . 7
- 2.4 Normalizing Test Speakers 8

3 Experiments 9
- 3.1 TIMIT Database and Training Conditions 10
- 3.2 Switchboard Database and Training Conditions 12

4 Results 12
- 4.1 Results on the TIMIT Corpus 13
- 4.2 Results on the Switchboard Corpus 13

5 Conclusions 25

1 Introduction

A complication in speaker independent automatic speech recognition compared to speaker dependent recognition is the increased acoustic variability due to speaker differences. A lot of the acoustic variability can be contributed to the differences in the length of the vocal tract. The average length of the vocal tract among males is significantly larger than the average among females. Therefore, some speaker normalization can be achieved by the use of gender dependent models but due to a significant variance in the vocal tract length within a gender group, considerable acoustic variability among speakers within a gender remains. More fine grained automatic normalization schemes were developed over the last couple of years. All these techniques incorporate vocal tract length normalization in the feature extraction phase. They differ however in the way the vocal tract length is estimated. Techniques for the estimation of the vocal tract can be divided in two groups First a *"knowledge based"* approach was developed by BBN[1], where vocal tract length differences are estimated from average formant locations. This approach was also investigated by others[2]. A second approach uses a likelihood based estimation of vocal tract length differences. In this approach the likelihood of a finite set of features, normalized for different vocal tract lengths given a speech model are computed. The vocal tract length is then estimated by determining which features generated the highest likelihood. In the initial work by Andreou *et. al.*[3] as well as later work by Dragon[4] and AT&T[5] the speech recognition system itself was used in the likelihood computation. To address the computational cost problem introduced in this way, Dragon developed an algorithm in which a text-independent multivariate mixture density was used for the likelihood computation[6]. As this approach showed comparable performance (an approximately 2% to 3% drop in word error rates) at much smaller computational cost, many other sites currently use this approach[5, 7, 2].

In this report, the background and implementation of speaker normalization using both the likelihood based vocal tract length estimation procedure as well as the formant based method are described. The implementation was done using the ASSM software package described in detail in[8]. For the implementation of the formant based method, the commercially available *XWaves+* package was used. In the likelihood approach, a 256 mixture zeroth order Polynomial Segment Model (PSM) was used to estimate likeli-

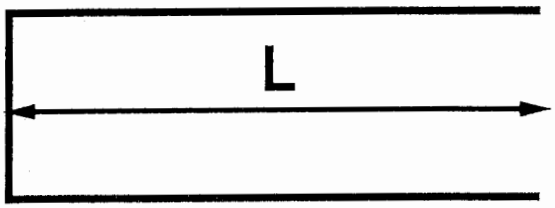

Figure 1: A simple tube model of the vocal tract

hoods. Section 2 describes the implemented algorithm, section 3 describes experimental conditions and section 4 describes the results obtained using the different normalization approaches.

2 Speaker Normalization Algorithms

In this section, the formant based and the maximum likelihood vocal tract length normalization (ML-VTLN) algorithm will be described. First in section 2.1 the algorithm for normalizing the features given a vocal tract length estimate is explained. Second in section 2.2 the formant based approach is described. Third in section 2.3 the algorithm used to train the mixture model for likelihood computations in the ML-VTLN framework is described. Finally, in section 2.4 the application of the normalization schemes for the recognition of test data (i.e. data not seen in training) is described.

2.1 Frequency warping

Given an estimate of the vocal tract length of a speaker relative to a mean vocal tract length (the derivation of this mean vocal tract length is described in section 2.2 for the formant based method and in section 2.3 for the ML-VTLN method), the spectral features derived from the speech waveform of that speaker are to be normalized such that acoustic differences due to a different vocal tract length are removed. To investigate the effect of the vocal tract length on the spectral features consider a simple tube model representing the vocal tract as shown in figure 2.1. The effect of changing the length L of the tube (i.e. varying the length of the vocal tract) will cause a linear shift in of the resonance frequencies k/L with $k \in \{1,3,\cdots\}$. Note though that the tube model is only a reasonable model for a schwa. A reasonable

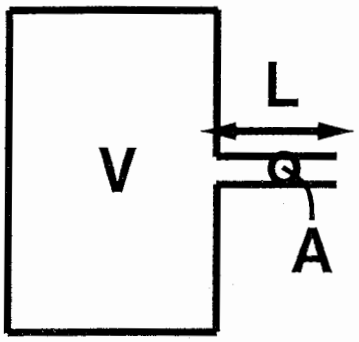

Figure 2: The Helmholz resonator model of the vocal tract

model for closed vowels such as /iy/ is the Helmholtz resonator depicted in figure 2 which has the first resonance frequencies dependent on the vocal tract length parameter L as $sqrt(V/AL)$ while the other resonance frequency dependencies are approximately linear. These simplified models show that in order to compensate for different vocal tract lengths a frequency warping can be used. The models also lead us to the conclusion that this frequency warping will have to be phone dependent. In this work however, each speaker will be limited to a single speaker dependent frequency warping to normalize spectral differences due to vocal tract length differences as is assumed in previous work by others. The warping function used is depicted in figure 3. The warping parameter α controls the slope of the linear warping from 0 to the fixed frequency Φ. From that point to the Nyquist frequency, the warping is also linear so as to reach the point $(1, 1)$. To implement this frequency warping the approach described by[6] is used, where the warped frequency axis is sampled at equally spaced intervals. For each warped frequency f' the corresponding original frequency f is computed. As the spectral representation is derived by an FFT, there is no guarantee that there is an estimate of the spectral energy at that exact frequency. To derive the spectral energy at arbitrary frequencies in between the discrete frequency estimates provided by the FFT, we use a linear interpolation of the spectrum estimated by the FFT. The warping process is depicted in figure 4

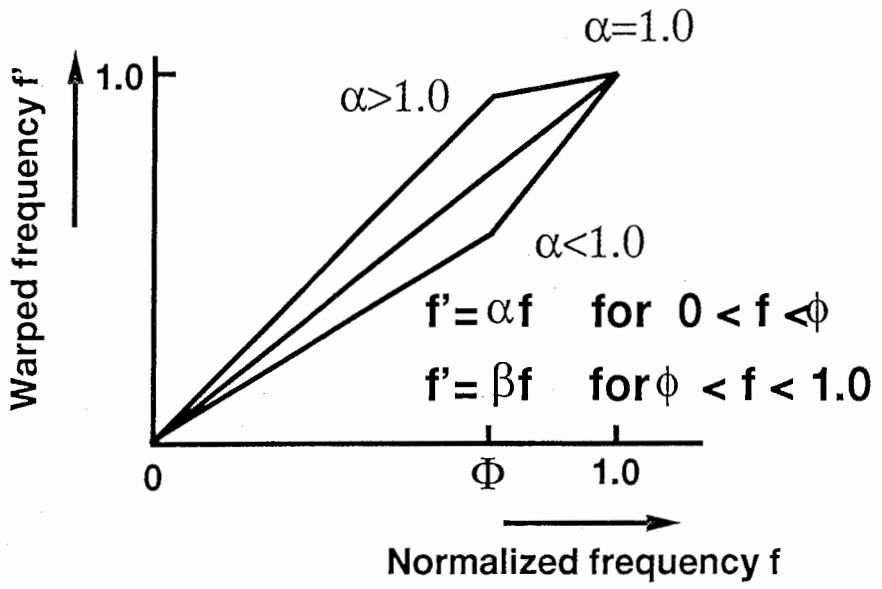

Figure 3: Piecewise linear frequency warping function

2.2 Formant based warp estimation

To determine the warping for a particular speaker, two related problems have to be solved. First, the "normalized" speaker has to be defined and second for each new speaker, the warping factor to normalize the speaker has to be determined. One approach is to use formant estimates to solve these problems[1]. The normalized speaker can be defined by computing the median formant location over all voiced frames from a training corpus. Then by computing the median formant location for a particular speaker by only using the voiced frames from that speaker, an estimate of the warping to normalize the features from that speaker is obtained. If the corpus median formant location is denoted as F_c and the median formant location of speaker S is denoted as F_S, the warping factor for that speaker is simply F_S/F_c.

The advantage of this approach is that continuous warp estimates are obtained. Another advantage is that this approach is possibly computationally inexpensive. A disadvantages of this technique are that the warping factors are sub-optimal in the maximum likelihood sense. Another disadvantage is that this technique can be computationally expensive when a very robust formant estimation algorithm is used. If a computationally inexpensive for-

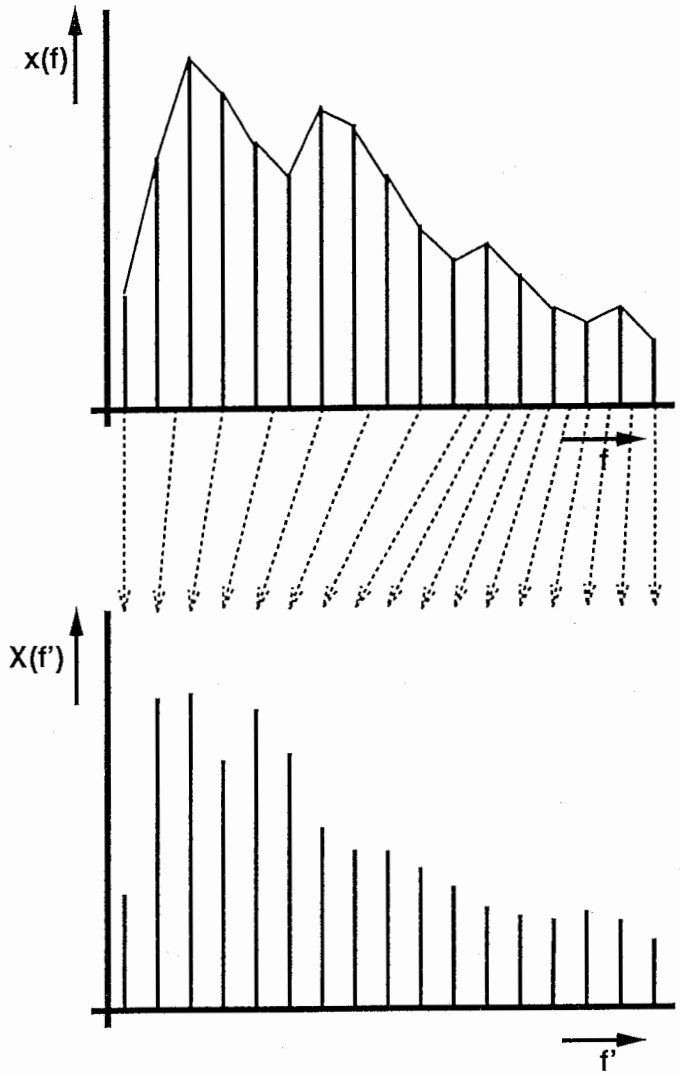

Figure 4: Equally spaced sampling of the warped frequency axis using spectrum estimation by linear interpolation

mant estimation algorithm is used it becomes more likely to get erroneous formant estimates and therefore erroneous warp estimates.

2.3 Maximum Likelihood Normalization using a Mixture Model

In the likelihood based approach to estimation of warping factors, an approach similar to[6] was used except that a segment based mixture model was used rather than a frame based one. To solve the related problems of defining the "normalized speaker" and to have an algorithm to automatically determine how to frequency-warp the data from an unknown speaker towards the normalized speaker, a likelihood based approach can be used. A text-independent segment based multivariate mixture model is trained for "the normalized" speaker. To train this model and simultaneously define the normalized speaker, the following training algorithm was used:

1. **Initialize:** Estimate a multivariate mixture model Λ_0 on unwarped features. Set $i = 0$;

2. **Likelihood estimation:** Compute the likelihood of the features of each training speaker $m = \{1, 2, \cdots, M\}$ warped at different warp factors $A \in \{\alpha_1, \alpha_2, \cdots, \alpha_N\}$ given the last model Λ_i.

3. **Estimate warps:** Estimate the most likely warp factor for each speaker m by determining which warping in A when applied to the features generated the highest likelihood given Λ_i. Let α_i^m denote the most likely warping for speaker m at iteration i.

4. **Retrain:** Retrain the mixture model using for each speaker the features warped at α_i^m.

5. **Iterate:** Set $i = i + 1$, go to 2.

A pictorial representation of this training algorithm is given in figure 5.

The mixture model was trained using a divise clustering approach in which each cluster is represented by a zeroth order PSM. The distance measure used in clustering is the negative log-likelihood of data with respect to the model parameters of a cluster[9].

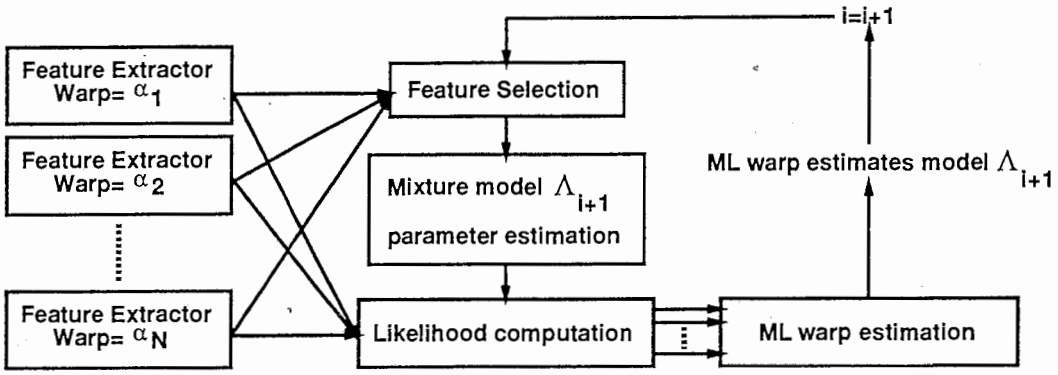

Figure 5: Model training overview

As the mixture model is implemented as a segment model, a segmentation has to be derived before this iterative training scheme can be executed. An acoustic segmentation algorithm is used to derive this segmentation. In this algorithm, described in detail in[9], stationary regions in the speech signal are sought by use of the dynamic programming algorithm. To derive a non-trivial segmentation, an average likelihood per frame threshold is used to control the average length of the stationary regions. Here it's important to prevent introducing a bias for certain warpings by allowing the segmentation for one warp factor to have a larger number of segments than the segmentation of the features at another warp factor. The segmentation is therefore derived in two steps. First the unwarped features are acoustically segmented. The the features at other warpings are segmented under the constraint that the number of segments per utterance is equal for each warping. Graphically, the segmentation is derived as shown in figure 6.

2.4 Normalizing Test Speakers

To normalize the features of a test speaker (i.e. a speaker not included in the training set), some or all of the speech available from that new new speaker is used to estimate the appropriate warping factor for that speaker. Typically 30 to 60 seconds of speech is used to estimate the warp factor of a speaker. After estimating the speaker dependent warp factor, all the features derived from the speech of that speaker are warped using this warping factor.

For the formant based approach to warp factor estimation the speech

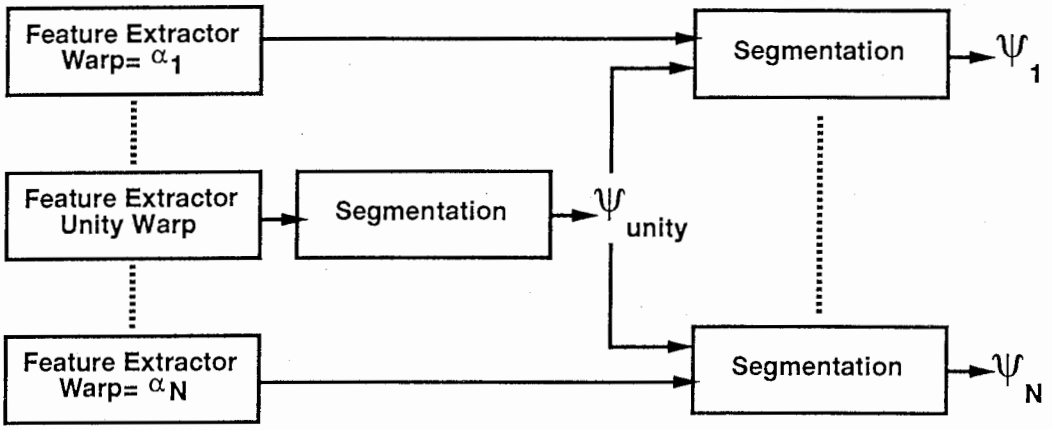

Figure 6: Segmentation overview

used for warp factor estimation is analyzed using the formant tracker used in the training process. Then the median formant location is computed which in comparison to the corpus median formant location gives an estimate for the speaker dependent warp factor.

In the likelihood based approach, features are extracted at all allowable warps from the speech used for warp factor estimation. The unwarped features are then acoustically segmented first. Subsequently, the features warped at the other factors are acoustically segmented under the constraint that the resulting segmentation should have the same number of segments as the unwarped segmentation does. The likelihood of these features given the trained likelihood model is then computed for all features at all warps. The speaker dependent warp factor is then determined by determining which warped features generated the highest likelihood. This warp factor estimation process is depicted in figure 7.

3 Experiments

Experiments were conducted on 2 corpora. Both the formant based warp factor estimation procedure and the likelihood based estimation procedure was used on the TIMIT read English corpus. The likelihood based method was also applied to the Switchboard spontaneous conversational English corpus. Evaluation of the effects of speaker normalization was performed on TIMIT

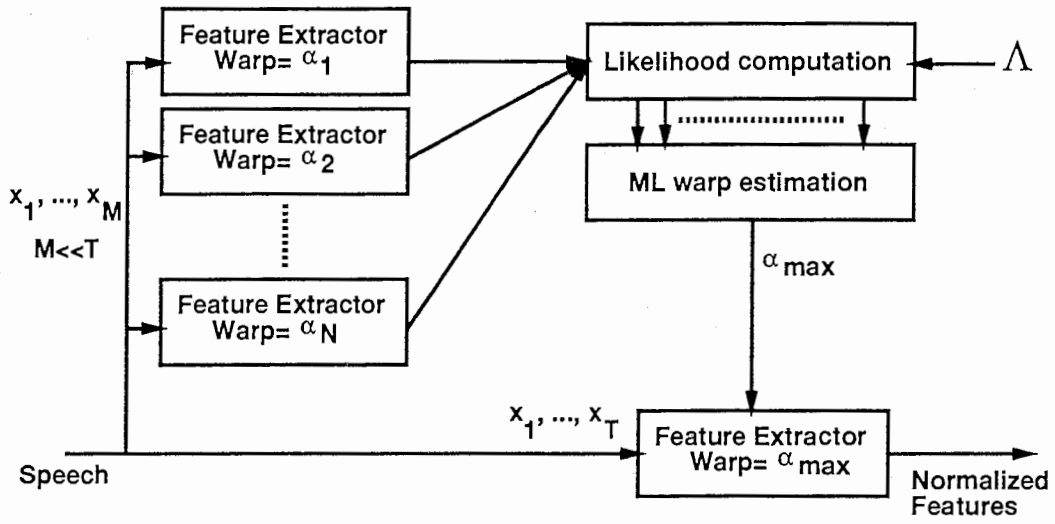

Figure 7: Test speaker warp factor estimation

by a phone classification experiment. Training and database information regarding the TIMIT corpus is given in section 3.1, regarding the Switchboard corpus in section 3.2.

3.1 TIMIT Database and Training Conditions

The TIMIT corpus consists of a training set of 462 speakers and a test set of 168 speakers with approximately 25 seconds of speech per speaker. The speech is read in a recording studio, digitized at a 16KHz sampling rate and quantized using 16 bits per sample.

The signal processing on the TIMIT corpus speech was performed using the following parameters:

- 25 ms. Hamming windowing
- 10 ms. frame shift
- $1 - 0.97z^{-1}$ pre-emphasis filter
- 4096 point fft
- 24 triangular filter filter-bank, equally spaced on the Mel scale

Formant	Median (Hz)
F1	557
F2	1534
F3	2597

- 14 dimensional Mel scale cepstral coefficient

The phone classification experiments on the TIMIT corpus were conducted using segment models with:

- 3 regions per model
- Single mixture Gaussian with constant mean and full covariance for each region.
- linear time warping

The standard Kai-Fu Lee phone set of 48 phones was used in the experiment. All of the training set defined in the TIMIT corpus, excluding the core sentences (sa1 and sa2) was used for training. Testing was done on all of the test set consisting of 50754 phones.

The formant based warp factor estimation was done using the first, second and third formant. Formant frequencies were estimated using the *XWaves+* software which uses the following algorithm:

1. Window waveform using a 25 ms Hamming window
2. derive AR model parameters by solving Yule-Walker equations
3. impose continuity constraint for formant tracks by dynamic programming

The corpus medians for the three formants were:

In the likelihood based TIMIT experiments, a gender balanced training set of 272 speakers, randomly selected from the available 462. A 256 mixture zeroth order mixture Polynomial Segment Model (PSM) was estimated. The set of allowable warp factors was $\alpha \in \{0.80, 0.82, \cdots, 1.20\}$.

3.2 Switchboard Database and Training Conditions

The Switchboard corpus is a conversational spontaneous speech corpus. The complete database consists of approximately 160 hours of telephone quality speech (digitized at 8kHz, quantized using 8bit mu-law coded samples).

For the experiments reported here, a gender balanced training set of 240 speakers, randomly selected from the available 2559 speakers was used. A total of 4 hours of speech was used for training, with approximately 60 seconds per speaker. Approximately 40 seconds of speech was used for the warp estimation of test speakers.

The signal processing on the Switchboard corpus speech was performed using the following parameters:

- 25 ms. Hamming windowing
- 10 ms. frame shift
- $1 - 0.97z^{-1}$ pre-emphasis filter
- 2048 point fft
- 24 triangular filter filter-bank, equally spaced on the Mel scale
- 14 dimensional Mel scale cepstral coefficient

For these experiments, both a gender independent (GI) and gender dependent (GD) models were trained. All models were 256 mixture zeroth order PSMs. The allowable warping parameters used in these experiments were $\alpha \in \{0.80, 0.82, \cdots, 1.36\}$. More warp factors were allowed in these experiments as other sites reported warp factor distributions beyond the range of warp factors used in the TIMIT experiments.

4 Results

The results of the experiments on the TIMIT corpus are given in section 4.1. The results on the Switchboard corpus are given in section 4.2.

Formant	Classification rate	Improvement
baseline	43.57%	-
F1	45.27%	1.70%
F2	45.24%	1.67%
F3	45.56%	1.99%

Table 1: Classification improvements due to speaker normalization

4.1 Results on the TIMIT Corpus

The warp factor distributions using the TIMIT warp factor estimation procedure using the first, second and third formant for both training and test sets are given in figure 8, 9 and 10 respectively.

The classification performance by use of speaker normalized features using the formant based approach are summarized in table 1.

The warp factor distributions using the ML-VTLN method on the TIMIT corpus is depicted in figures 11 through 15 for 5 training iterations. The classification results and data likelihoods using the warp factor estimation model obtained after each iteration is depicted in figure 16

Note that the likelihood increase of the training data going from the forth to the fifth iteration is very small and that the likelihood of the test data decreases at this iteration. The classification result shows a similar trend.

4.2 Results on the Switchboard Corpus

The warp factor distributions of the speaker in the training set during the 5 iterations of training of the GI model are depicted in figures 17 through 21. The warp factor distributions of the speaker in the training set during the 5 iterations of training of the male GD model are depicted in figures 22 through 26. The warp factor distributions of the speaker in the training set during the 5 iterations of training of the female GD model are depicted in figures 27 through 31. The distribution of warp factors for all the 2559 speakers in the corpus using the GD models are depicted in figures 32 and 33 for males and females respectively. The data likelihood given the models at different iteration steps is given in figure 34.

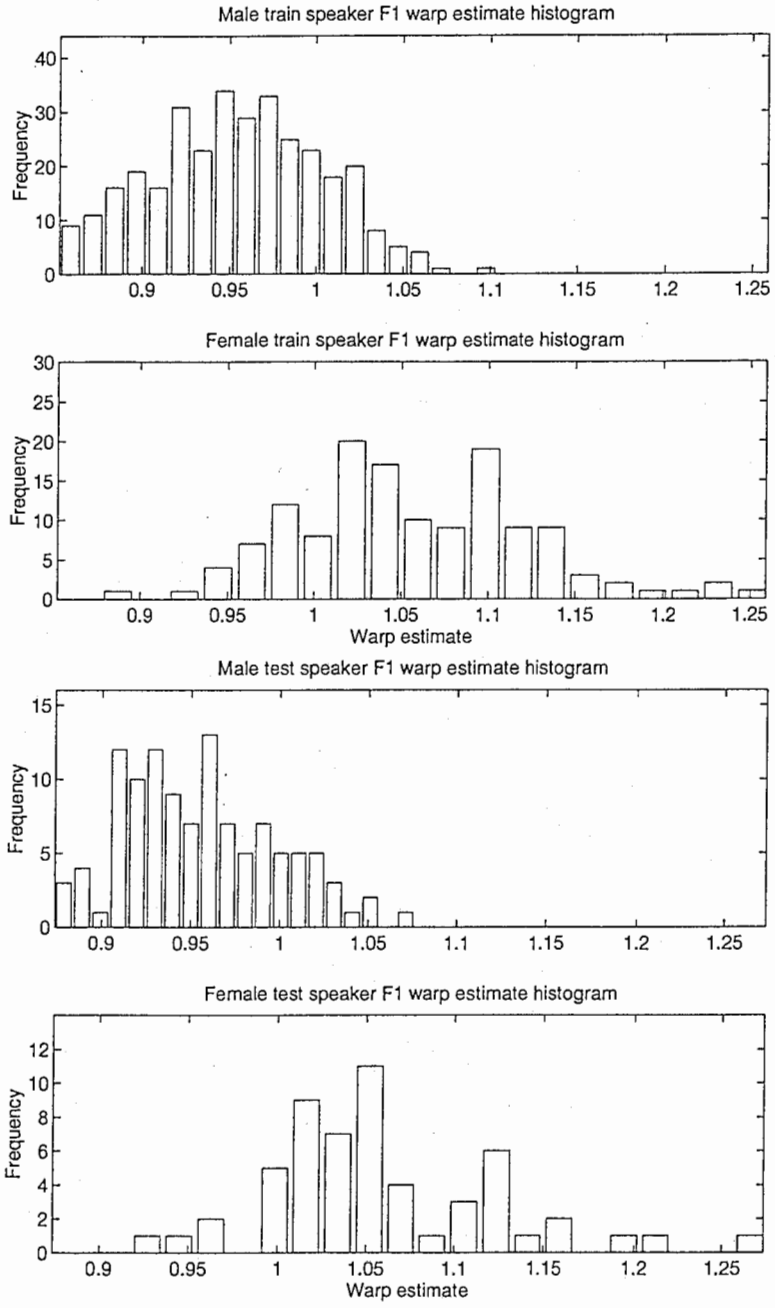

Figure 8: F1 based warp factor histograms for training and test sets

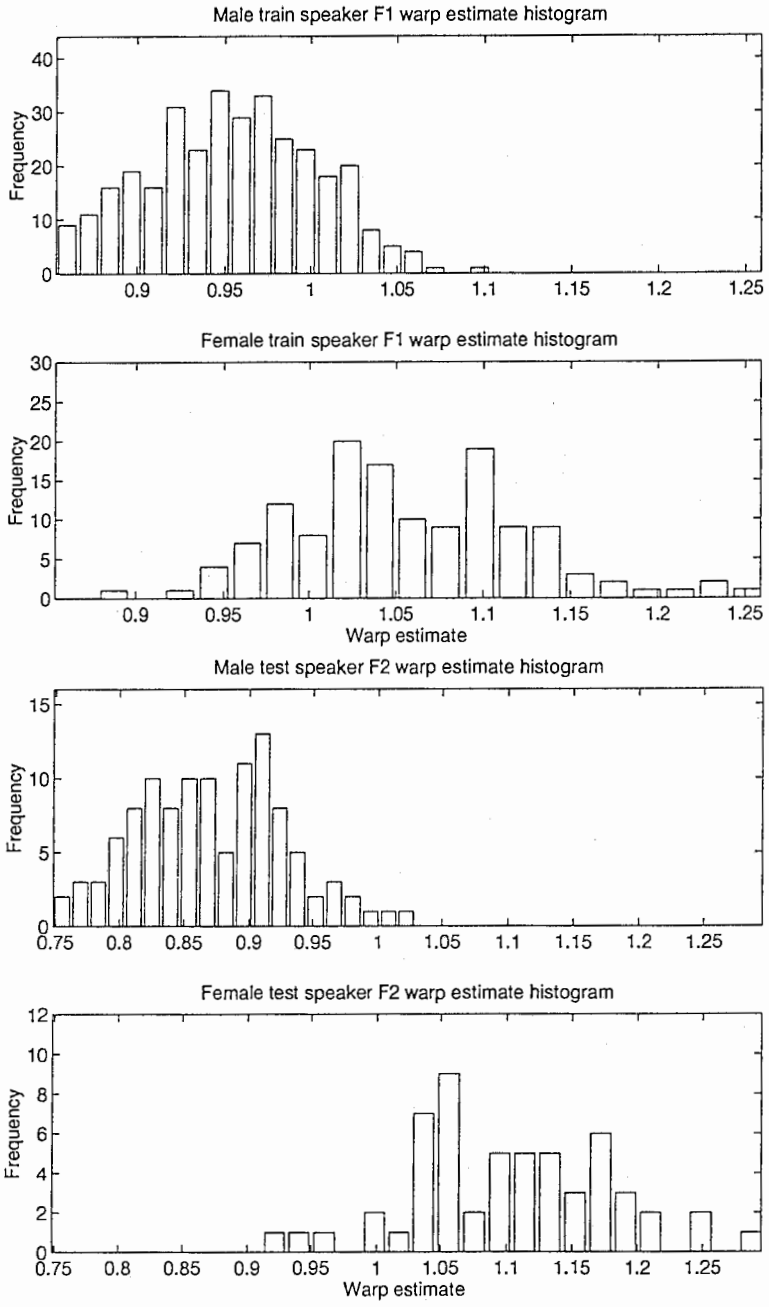

Figure 9: F2 based warp factor histograms for training and test sets

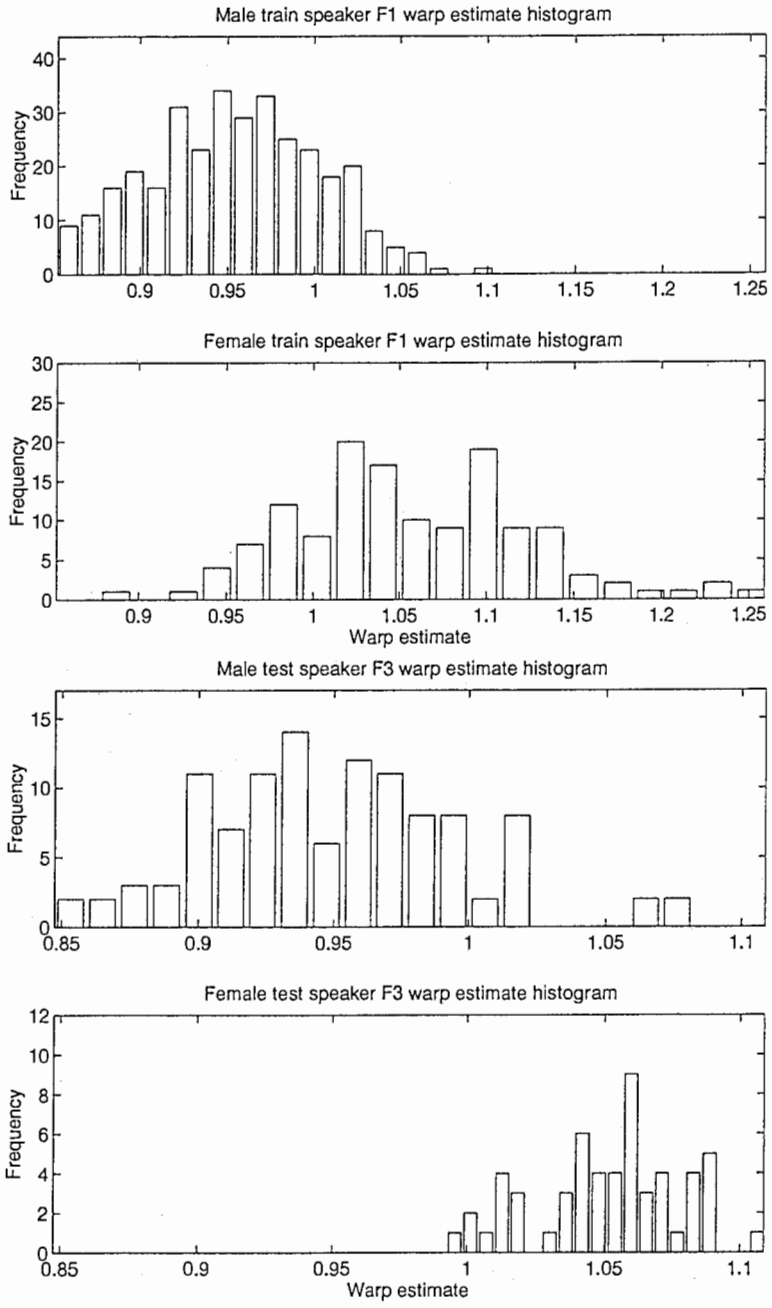

Figure 10: F3 based warp factor histograms for training and test sets

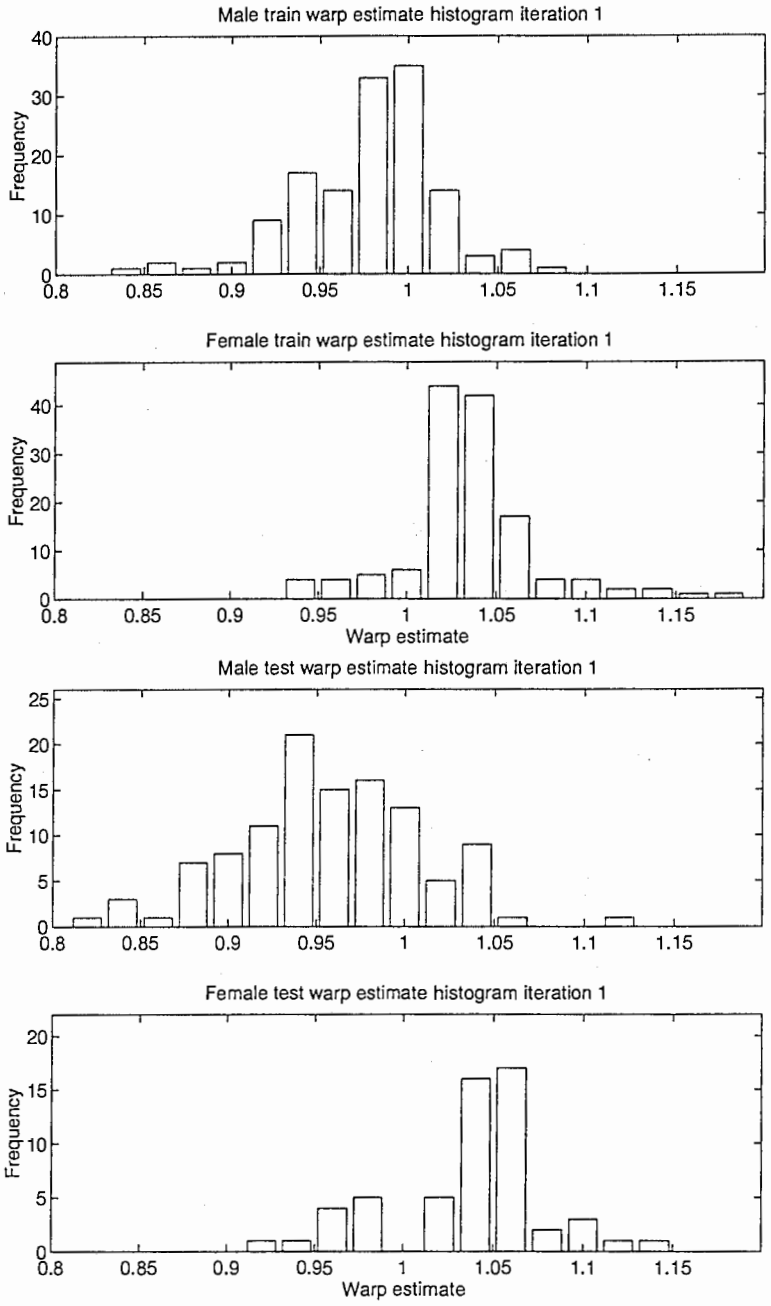

Figure 11: Warp factor histogram of training and test data using the mixture model estimated after iteration 1

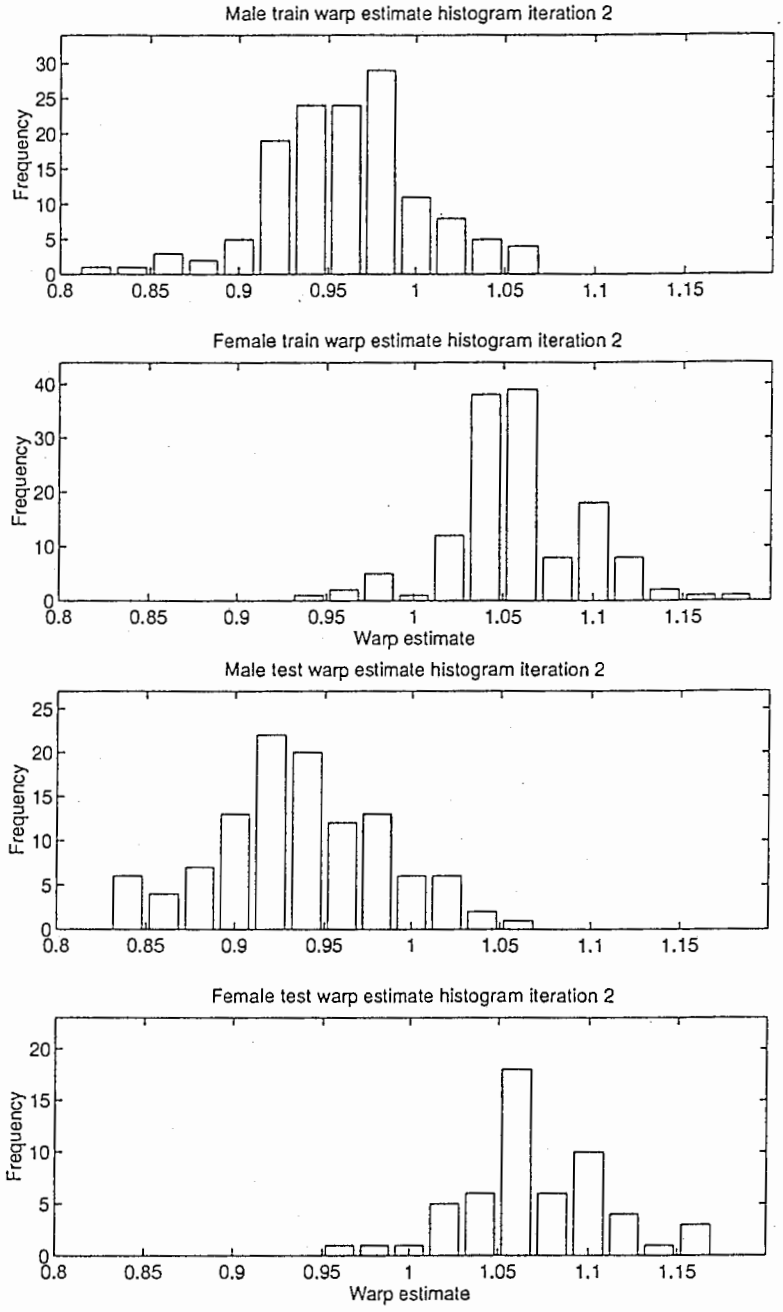

Figure 12: Warp factor histogram of training and test data using the mixture model estimated after iteration 2

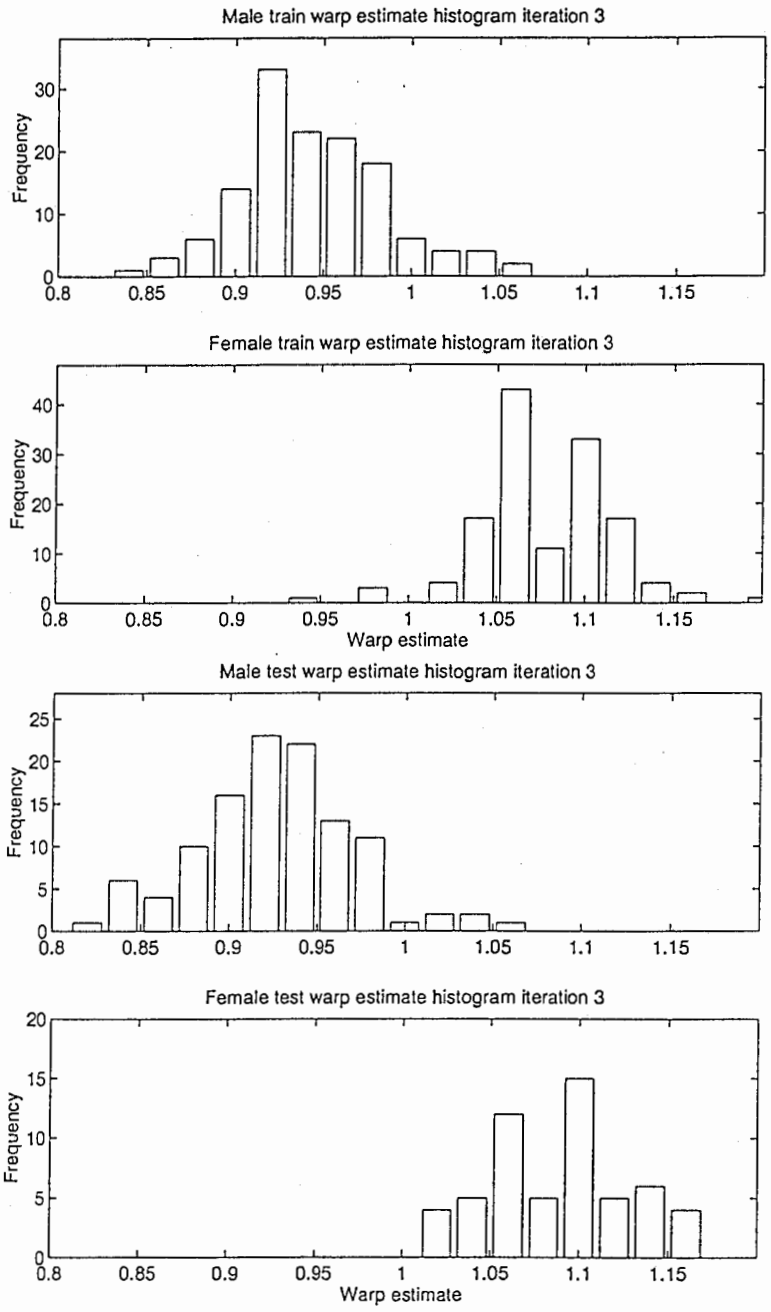

Figure 13: Warp factor histogram of training and test data using the mixture model estimated after iteration 3

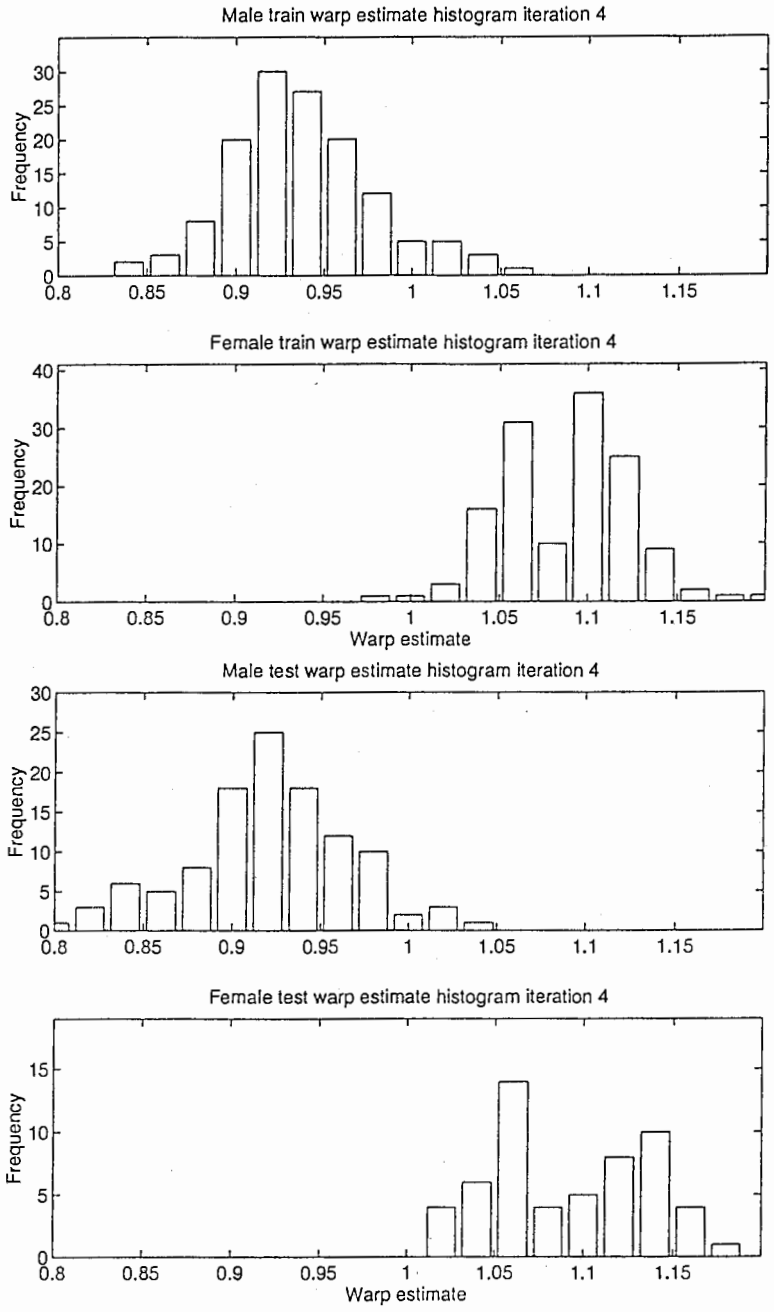

Figure 14: Warp factor histogram of training and test data using the mixture model estimated after iteration 4

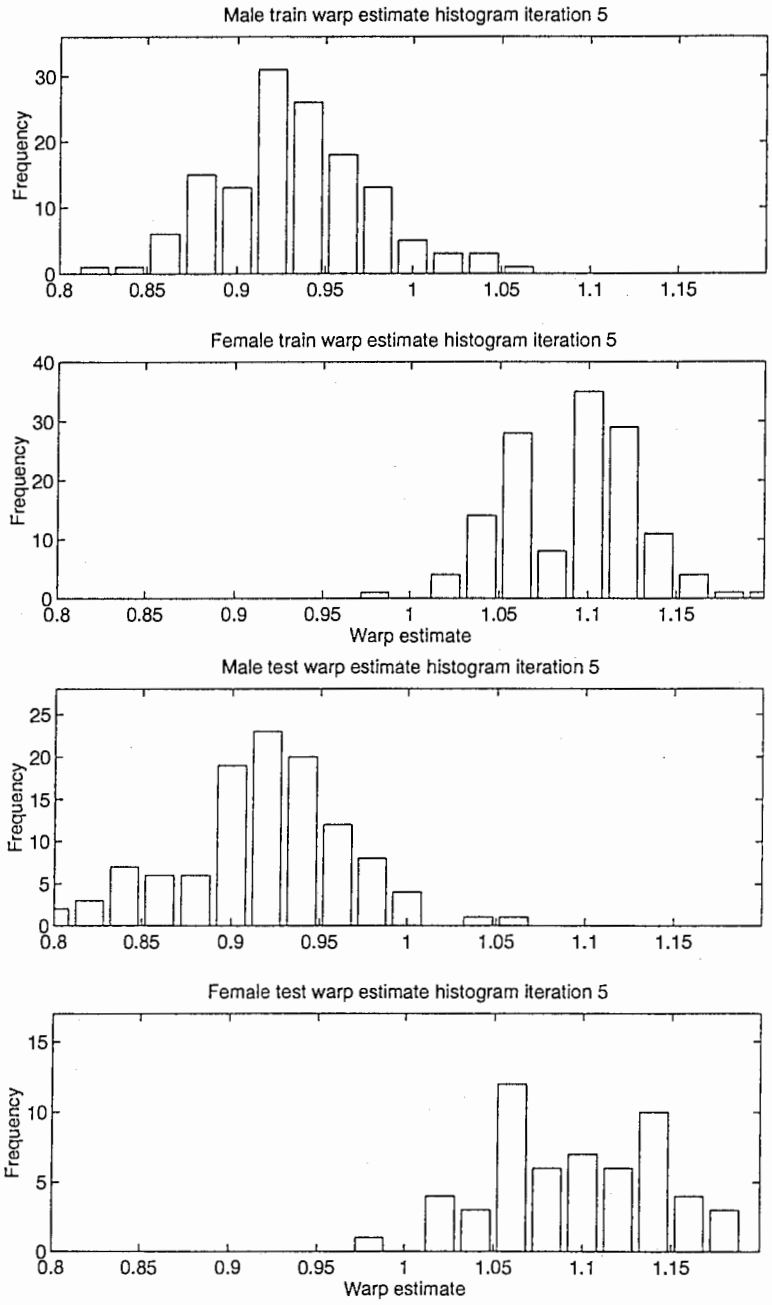

Figure 15: Warp factor histogram of training and test data using the mixture model estimated after iteration 5

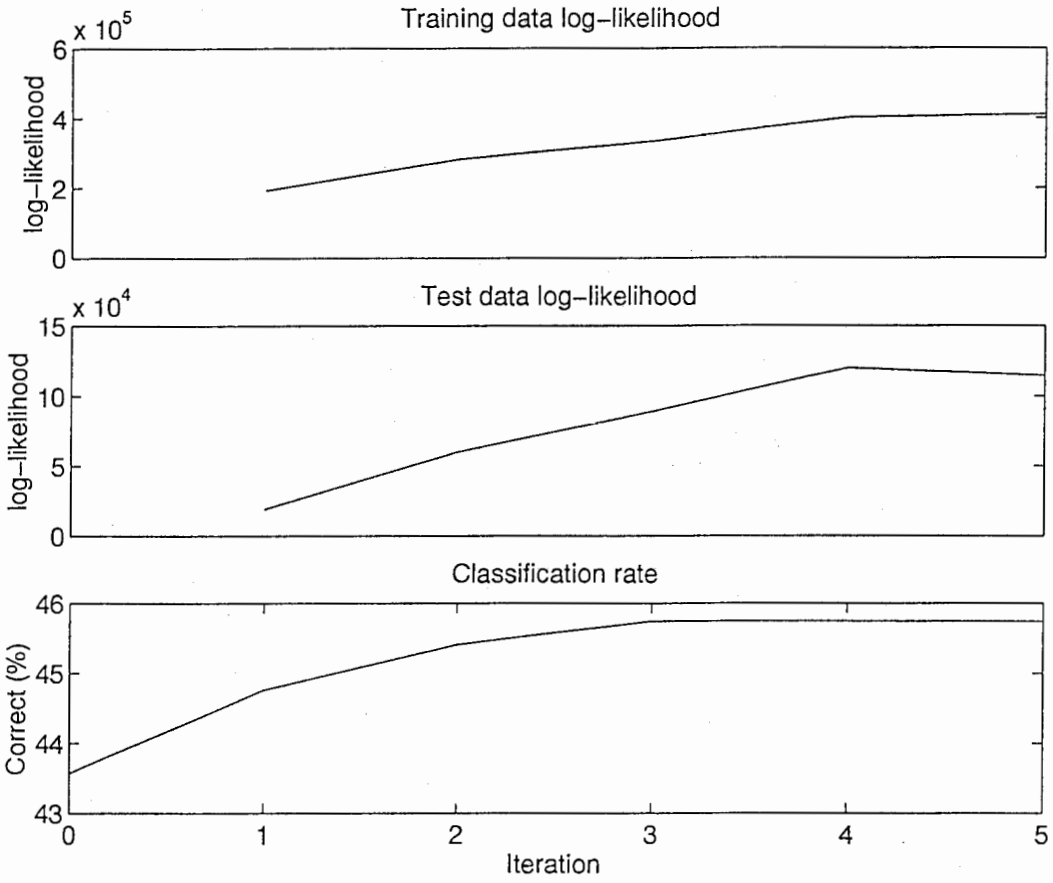

Figure 16: Classification scores and data likelihoods using the different mixture models estimated after each training iteration step

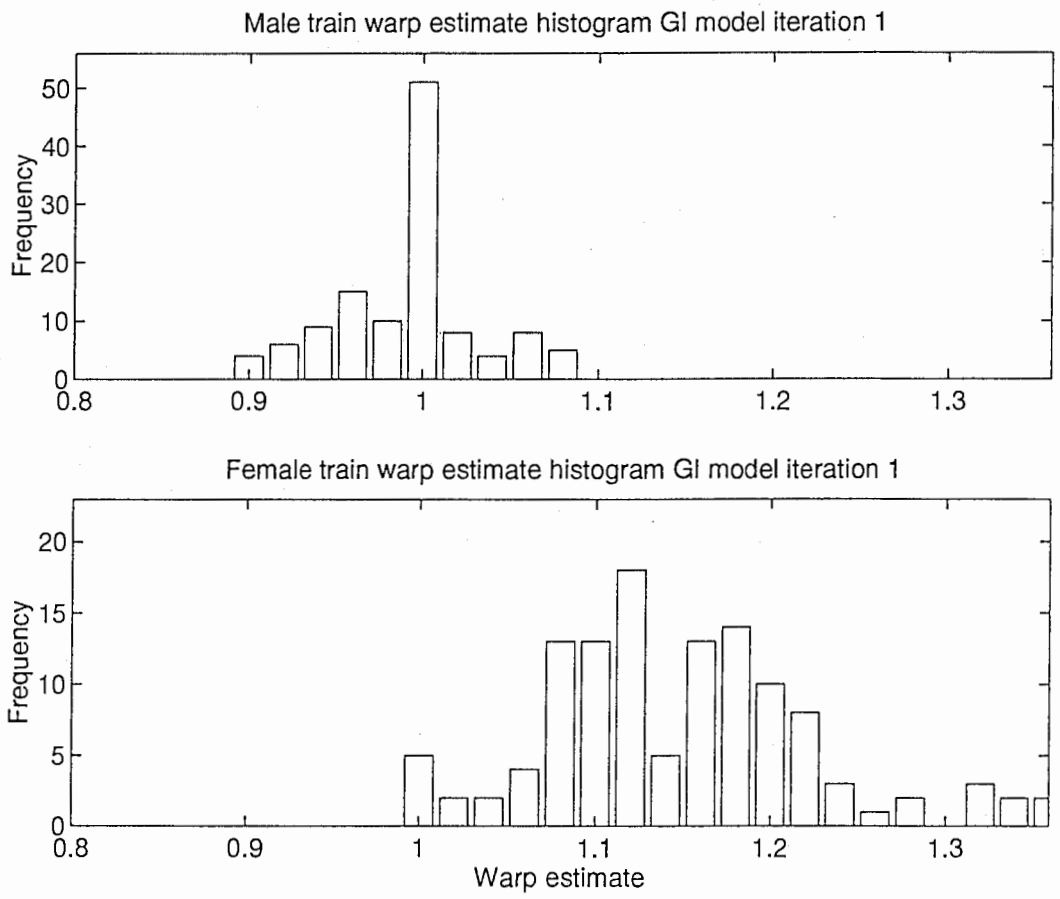

Figure 17: Warp factor histogram of training data using the gender independent mixture model estimated after iteration 1

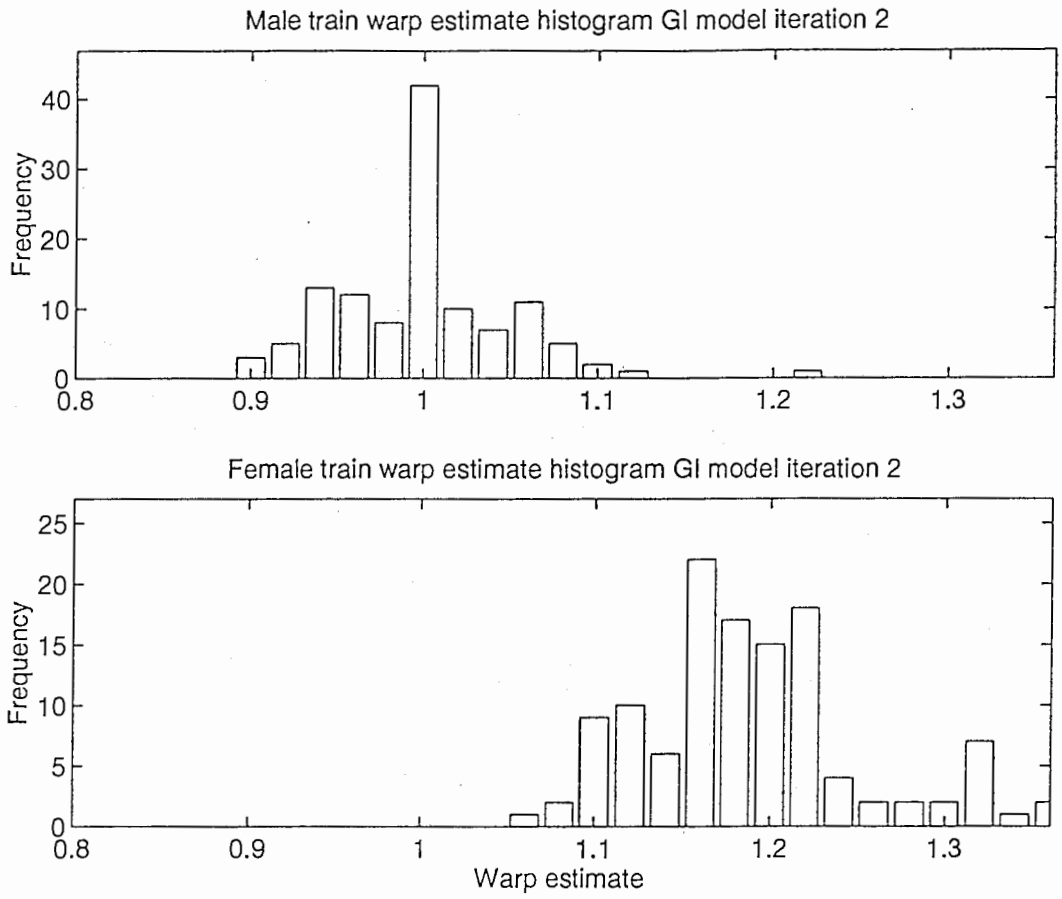

Figure 18: Warp factor histogram of training data using the gender independent mixture model estimated after iteration 2

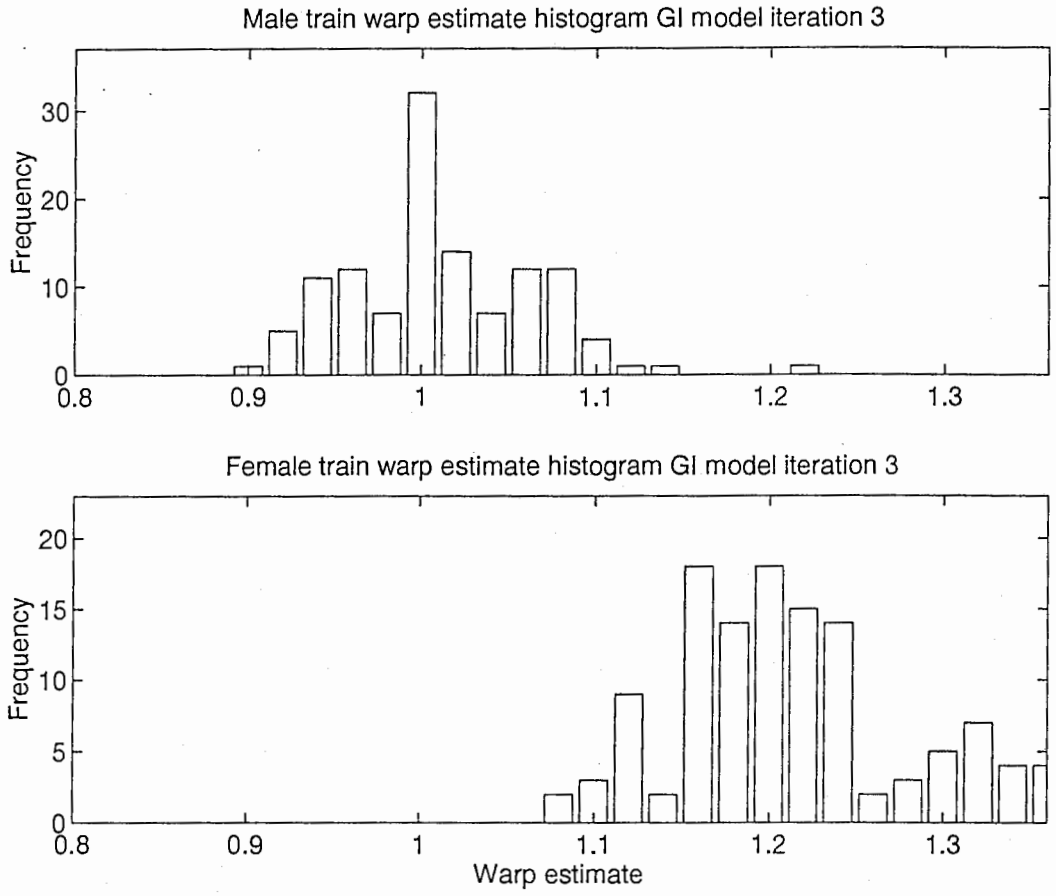

Figure 19: Warp factor histogram of training data using the gender independent mixture model estimated after iteration 3

5 Conclusions

The speaker normalization using either the formant based warp estimation approach as well as the likelihood based approach both give gains in classification performance on the TIMIT corpus similar to the gains reported by other sites.

For both corpora, the gender independent models show a clear separation of the distributions of warp factors of males and females but also show a considerable spread around the gender dependent mean warp factor.

It can be noted that during the iterative training process on the Switch-

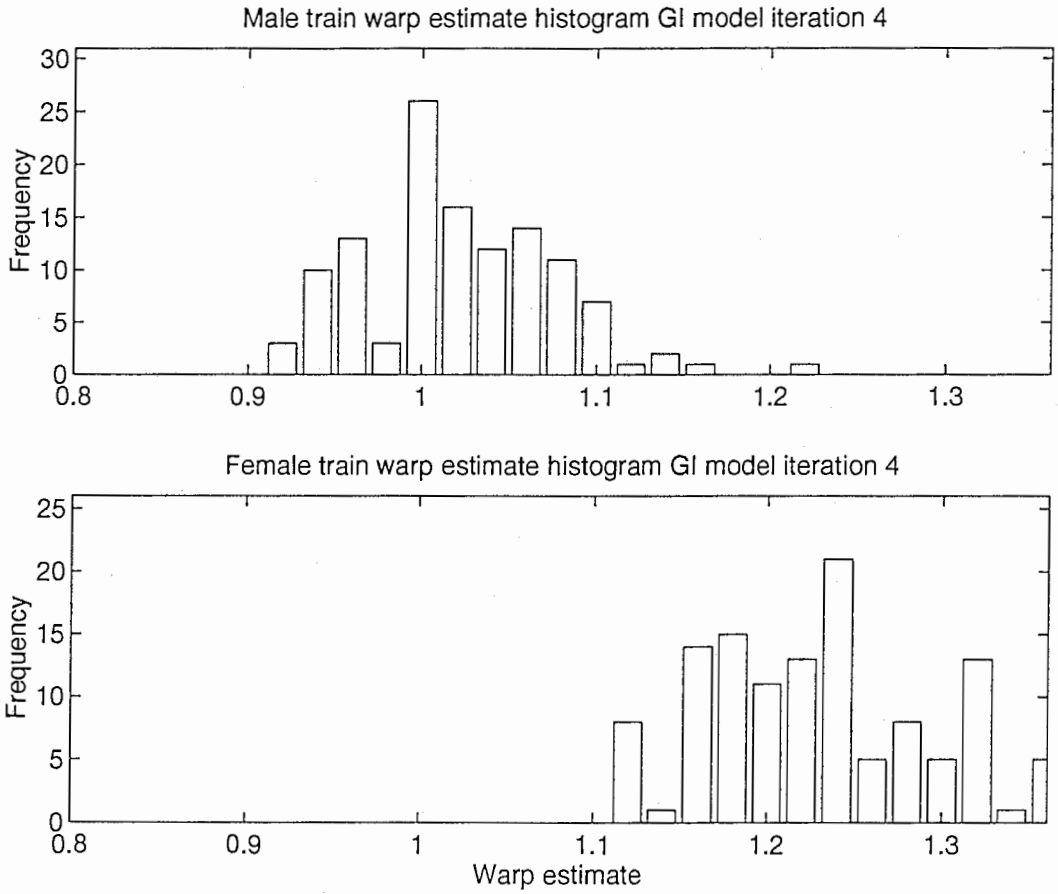

Figure 20: Warp factor histogram of training data using the gender independent mixture model estimated after iteration 4

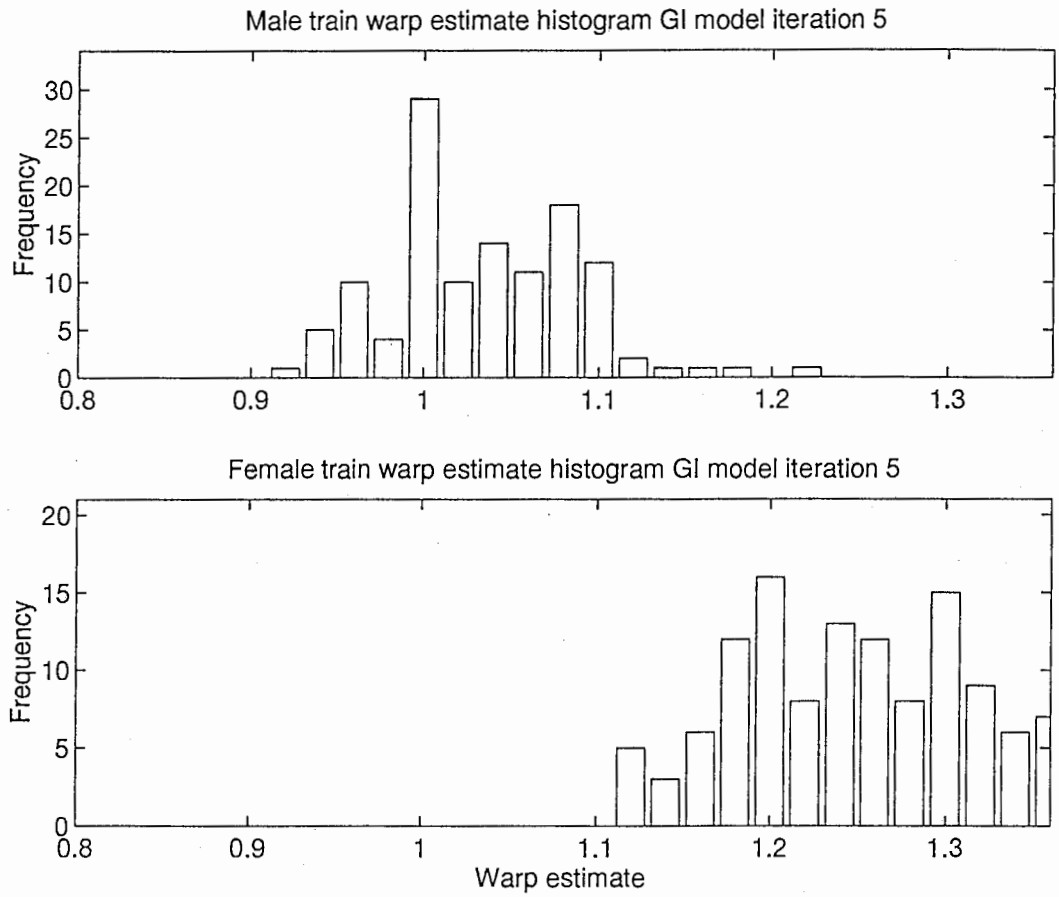

Figure 21: Warp factor histogram of training data using the gender independent mixture model estimated after iteration 5

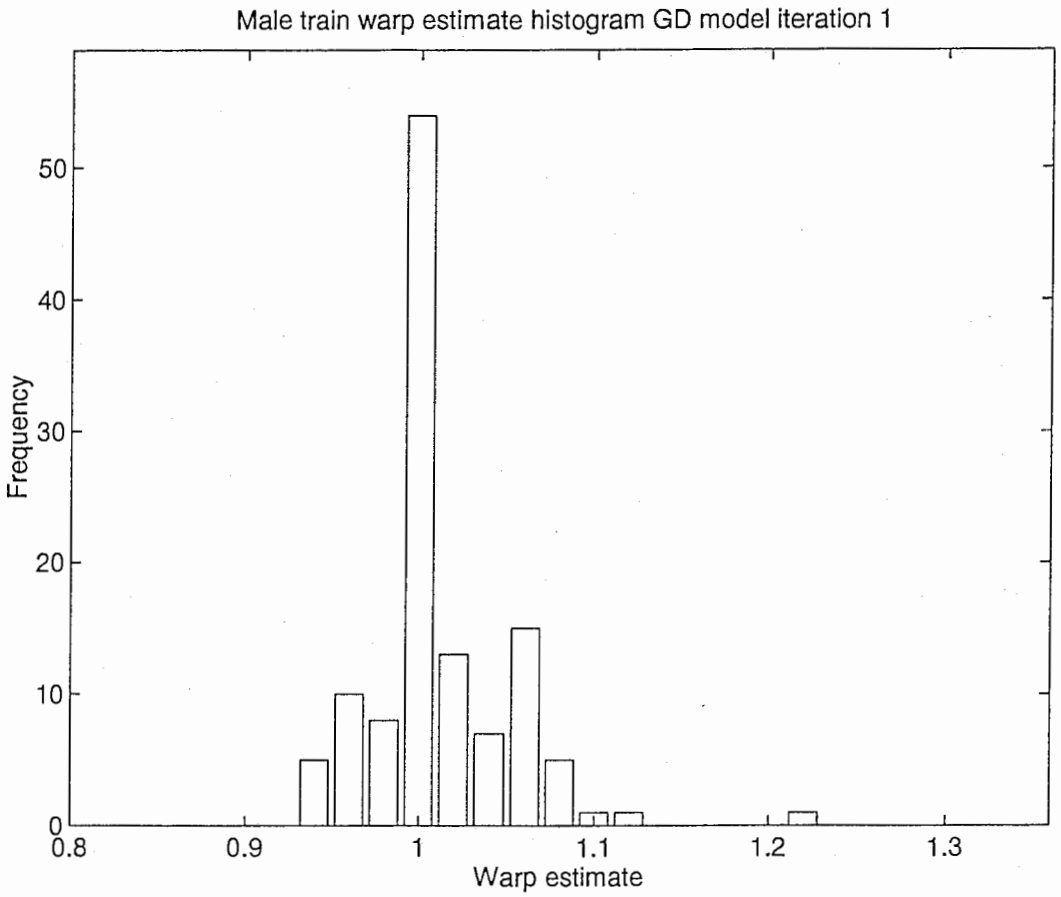

Figure 22: Warp factor histogram of training data using the male gender dependent mixture model estimated after iteration 1

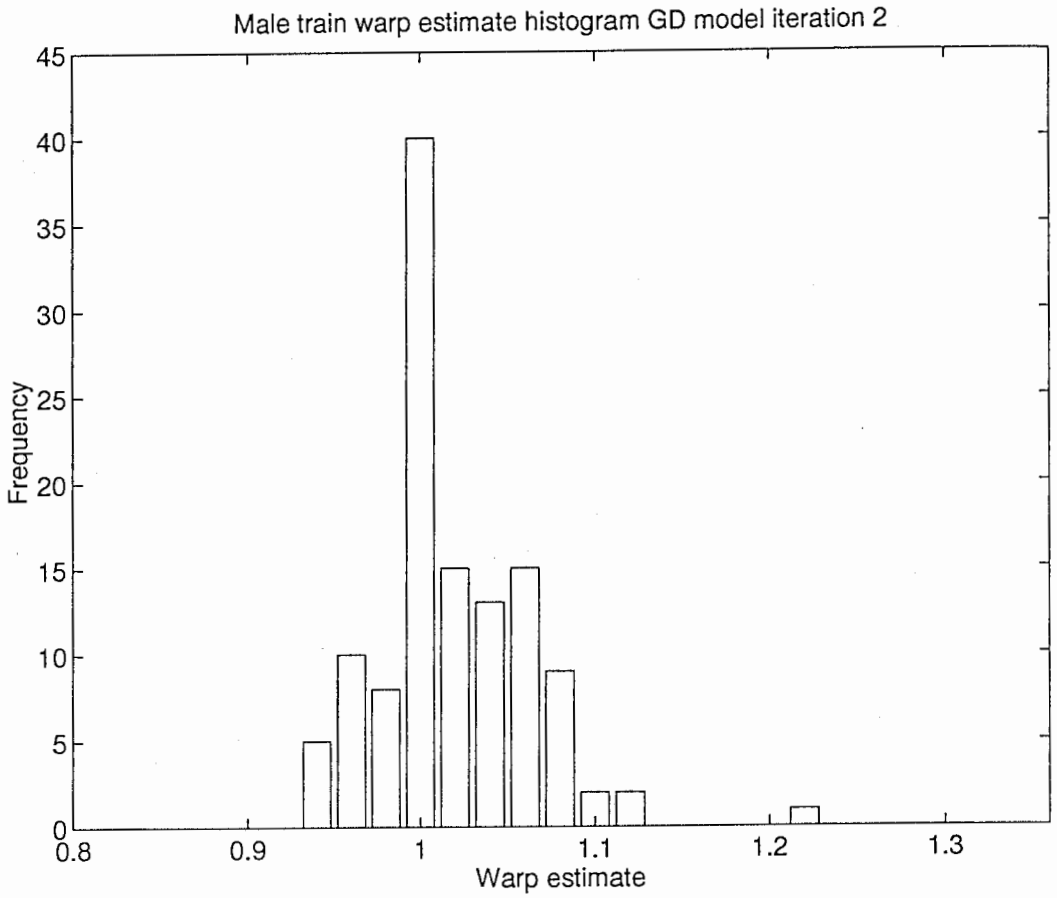

Figure 23: Warp factor histogram of training data using the male gender dependent mixture model estimated after iteration 2

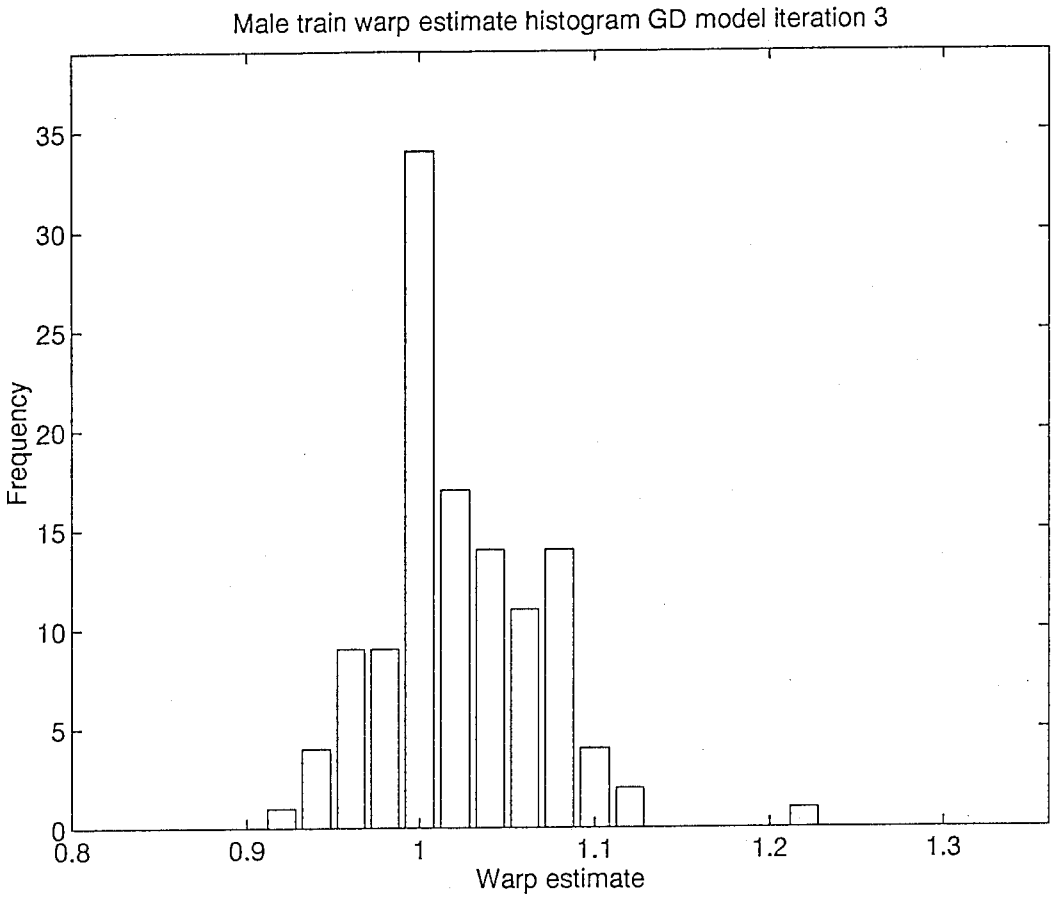

Figure 24: Warp factor histogram of training data using the male gender dependent mixture model estimated after iteration 3

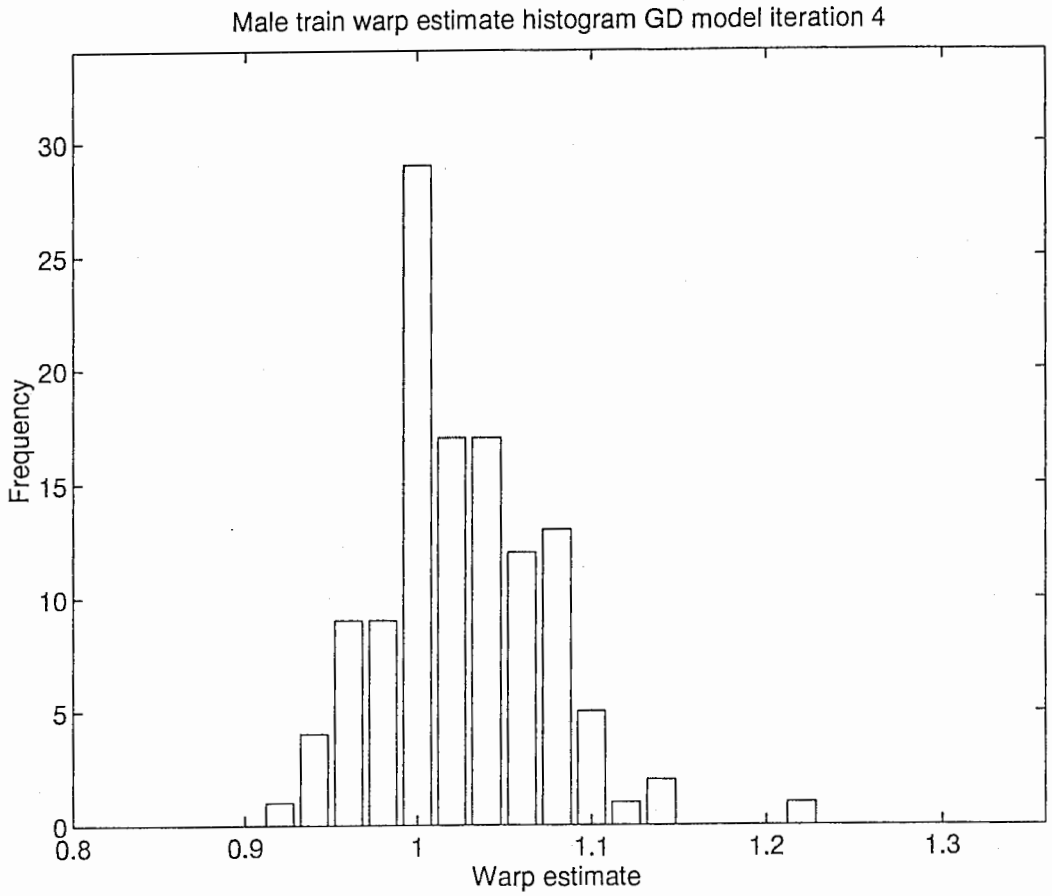

Figure 25: Warp factor histogram of training data using the male gender dependent mixture model estimated after iteration 4

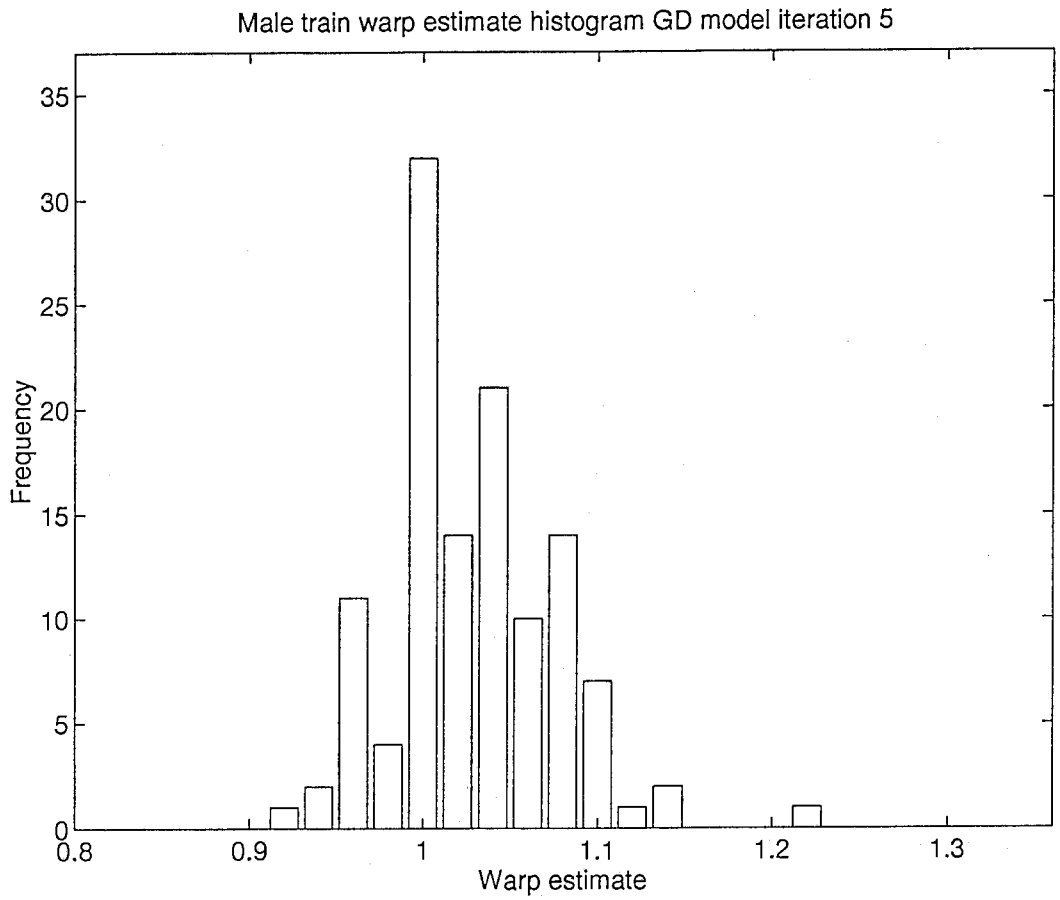

Figure 26: Warp factor histogram of training data using the male gender dependent mixture model estimated after iteration 5

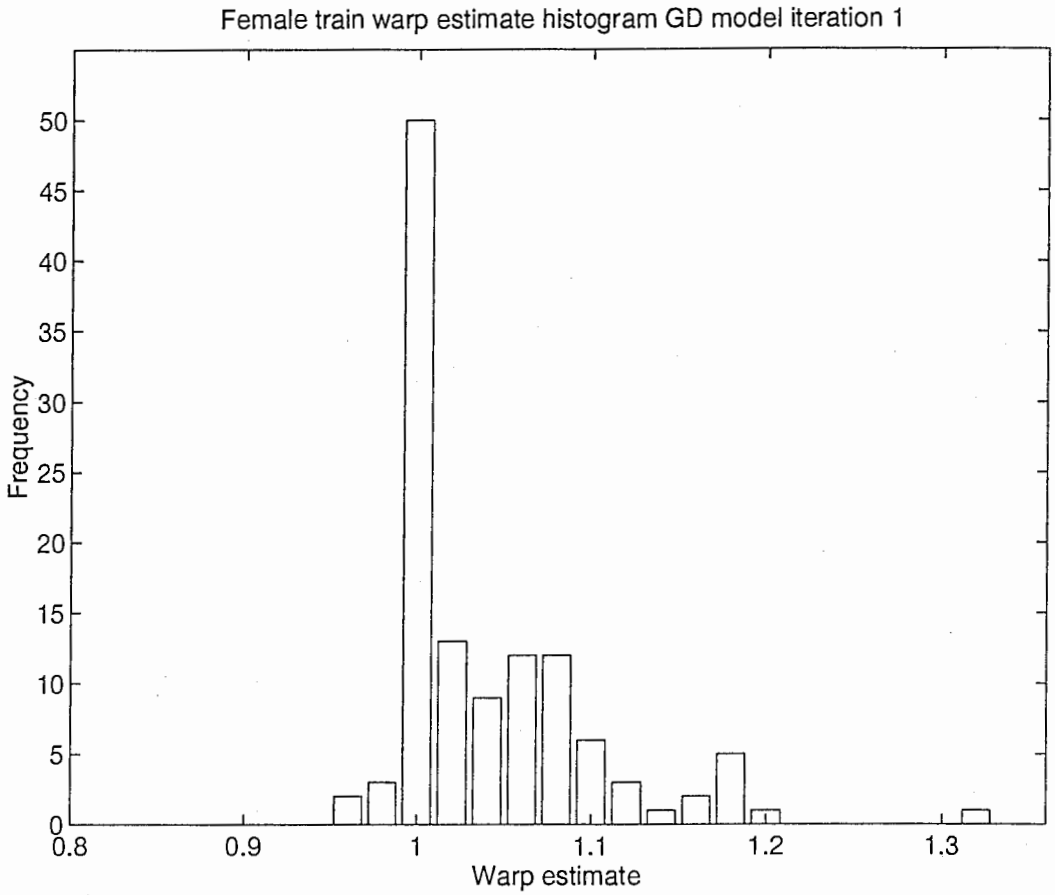

Figure 27: Warp factor histogram of training data using the female gender dependent mixture model estimated after iteration 1

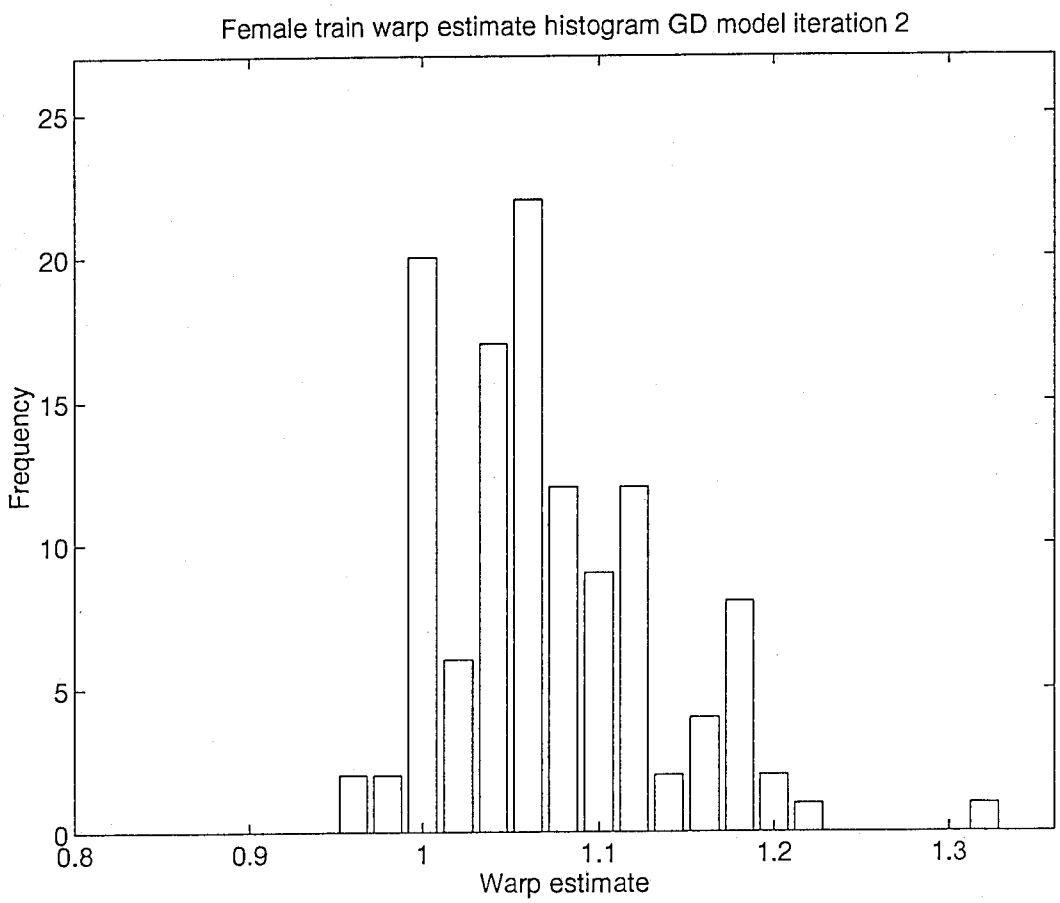

Figure 28: Warp factor histogram of training data using the female gender dependent mixture model estimated after iteration 2

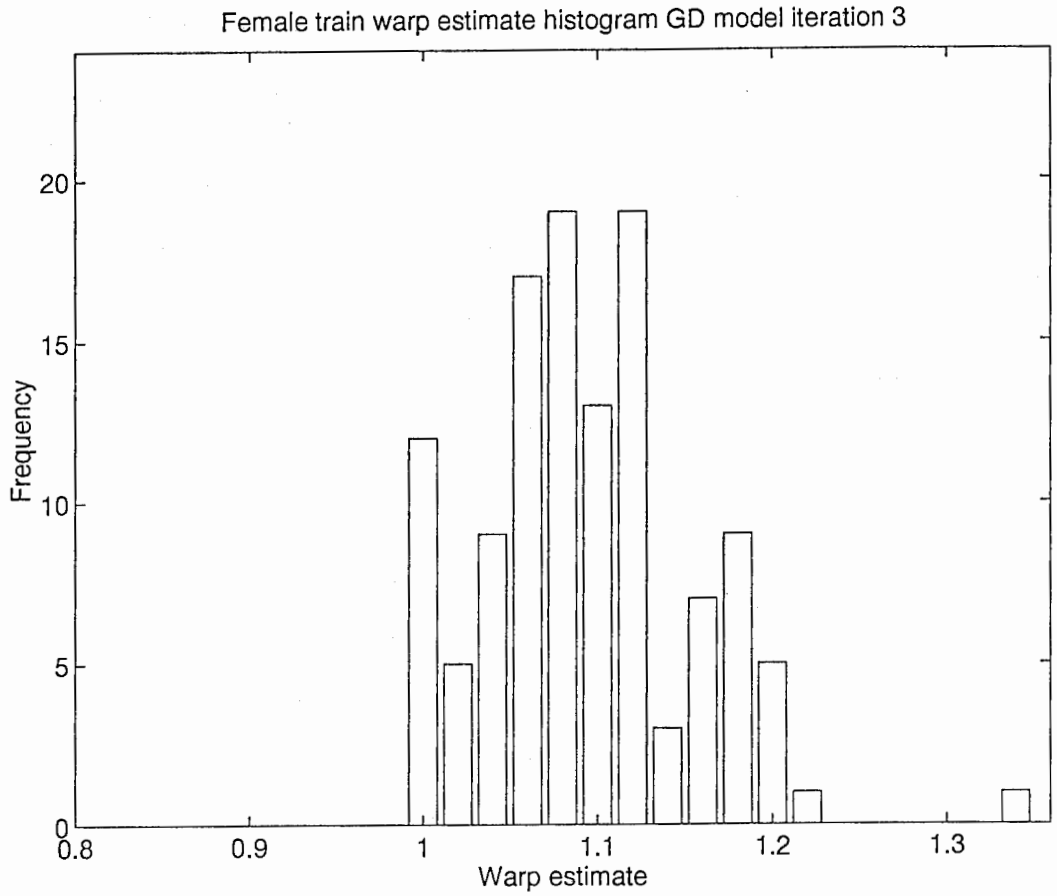

Figure 29: Warp factor histogram of training data using the female gender dependent mixture model estimated after iteration 3

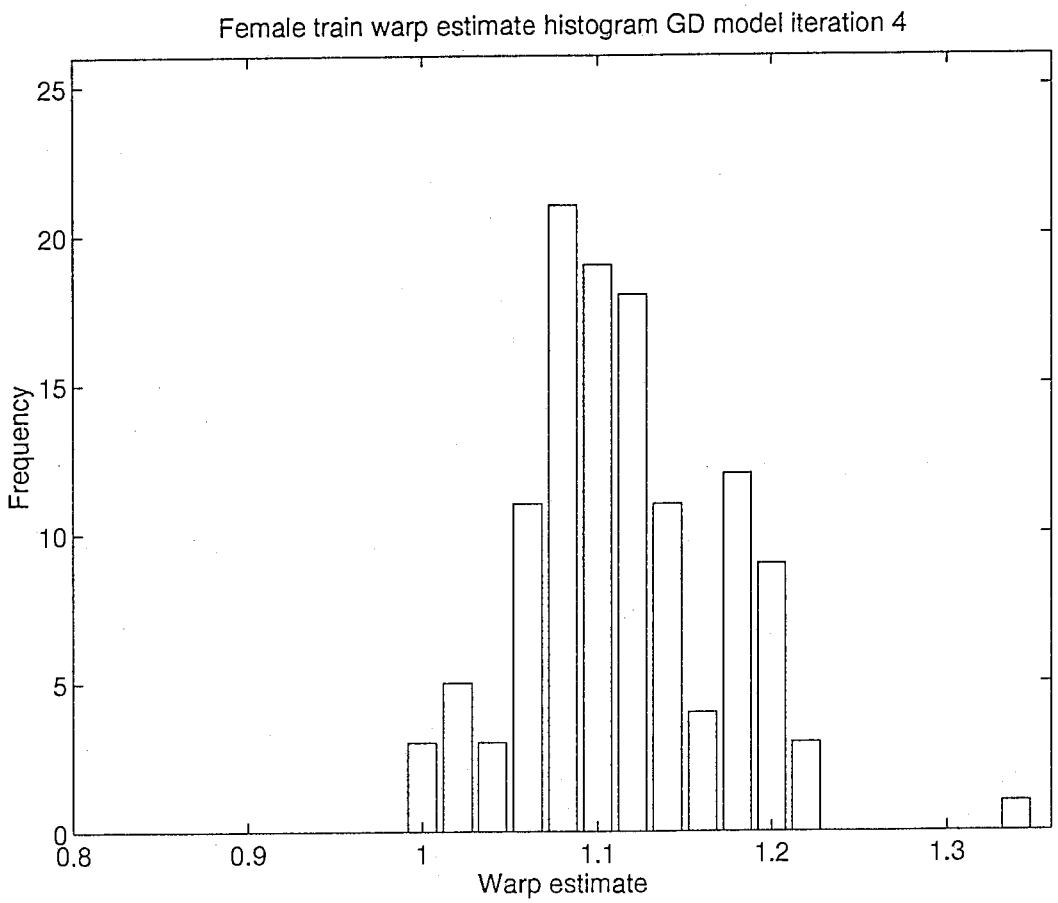

Figure 30: Warp factor histogram of training data using the female gender dependent mixture model estimated after iteration 4

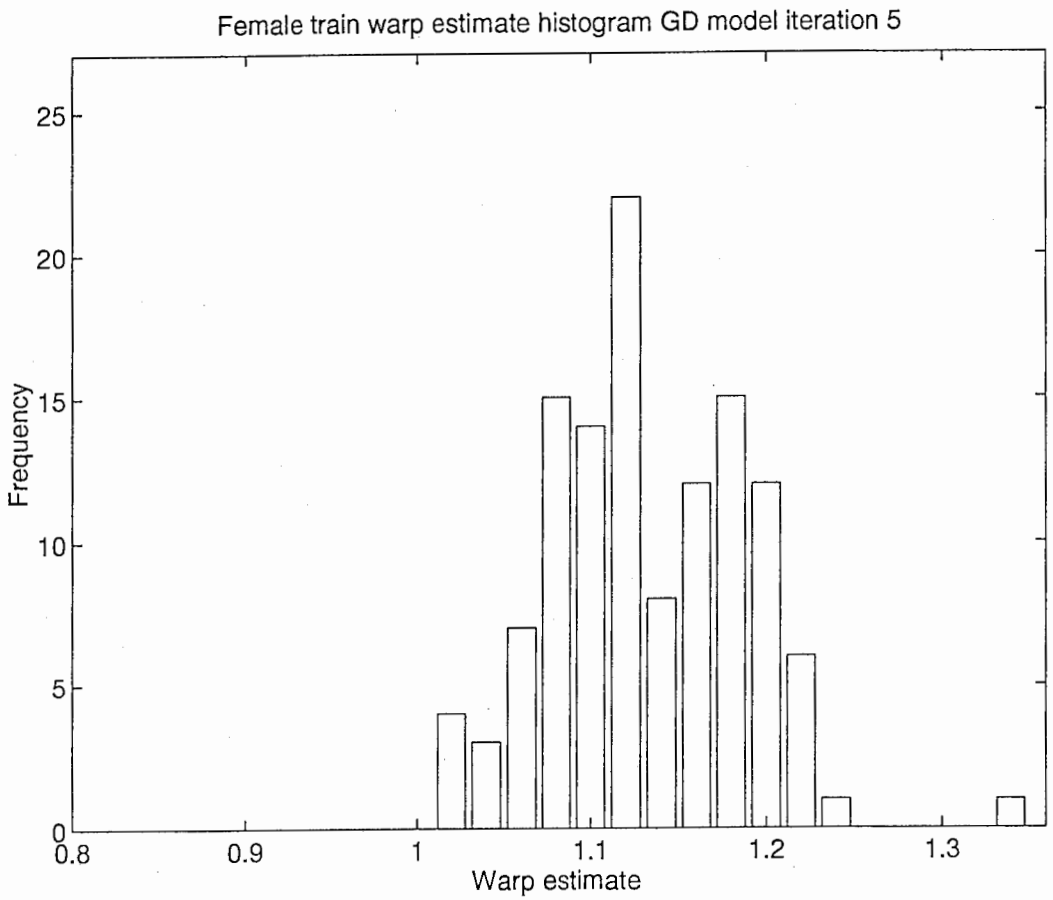

Figure 31: Warp factor histogram of training data using the female gender dependent mixture model estimated after iteration 5

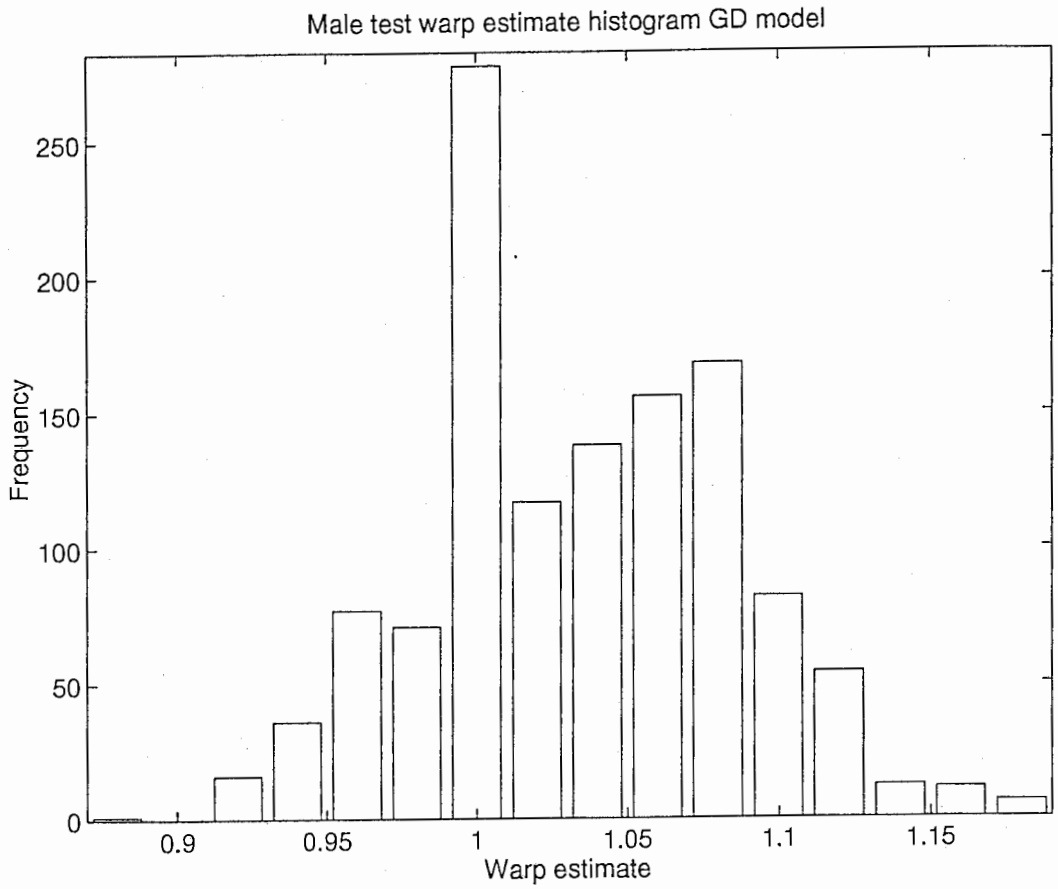

Figure 32: Warp factor histogram of test data using the male gender dependent mixture model estimated after iteration 4

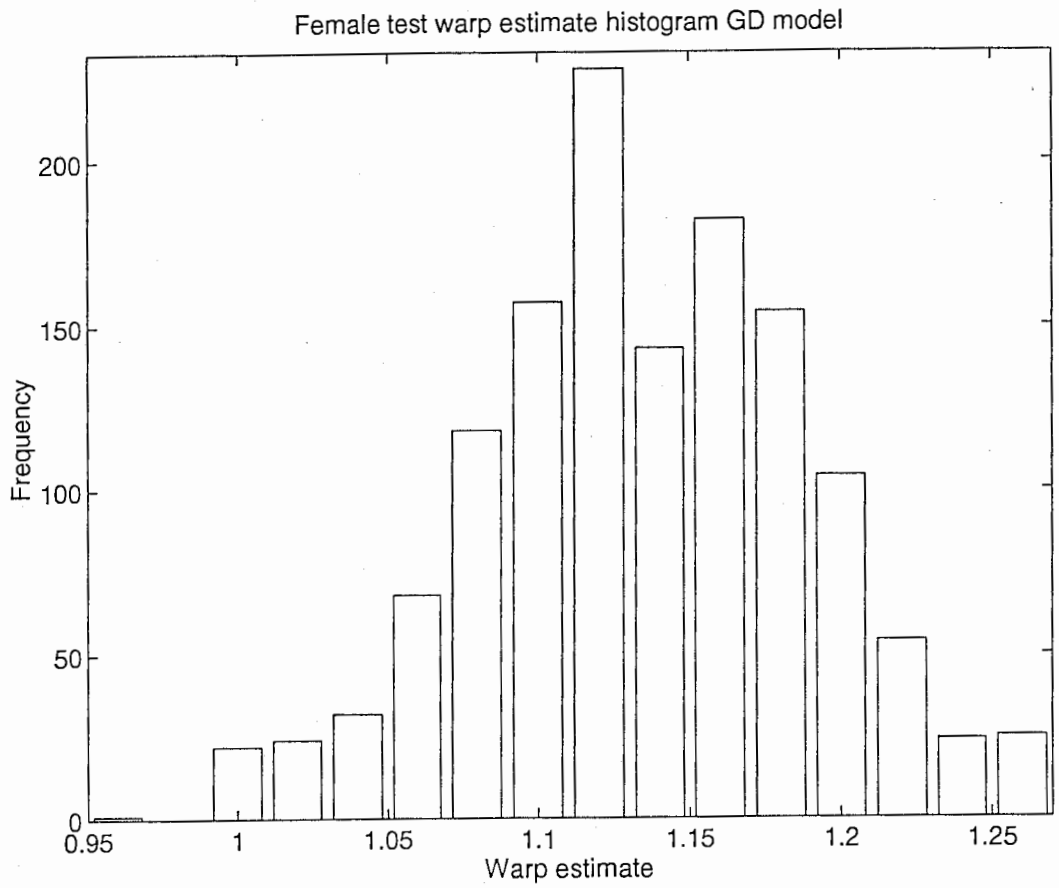

Figure 33: Warp factor histogram of test data using the female gender dependent mixture model estimated after iteration 4

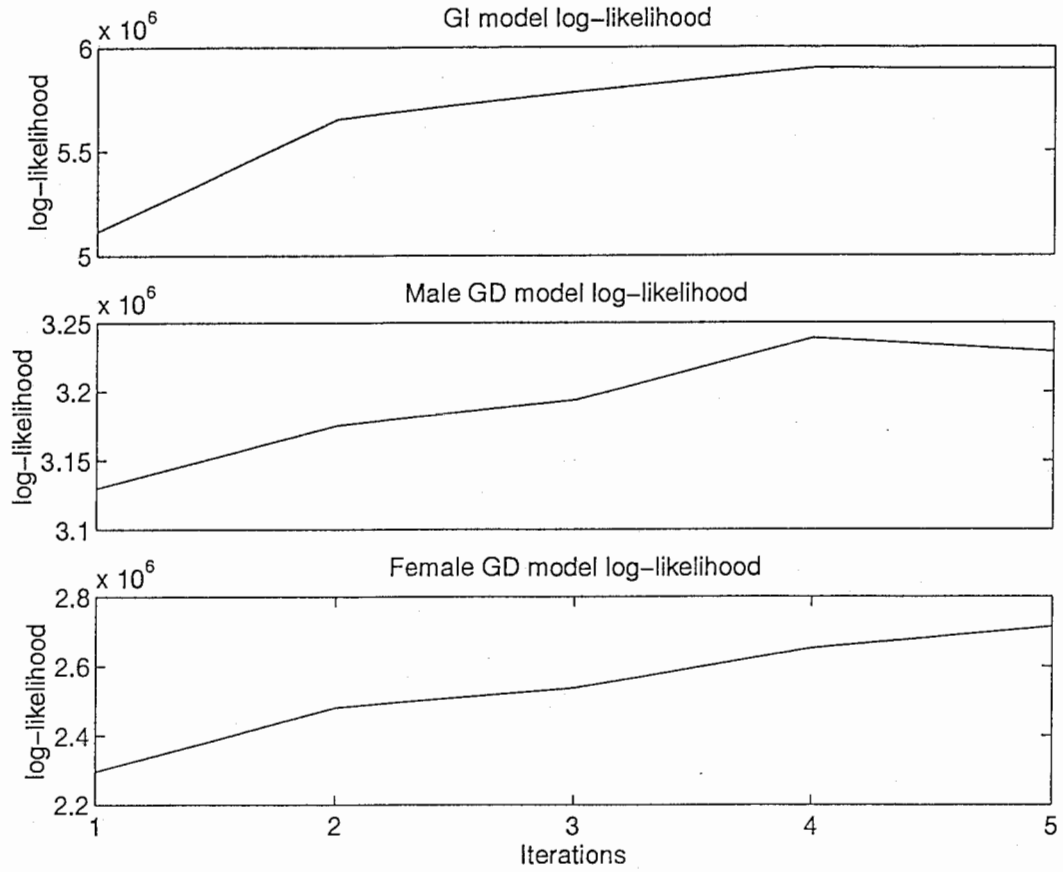

Figure 34: Data likelihoods of the training data given the model estimated after the different training iterations

board corpus, the median of the histogram of warp factors drifts towards higher warp factors. This was independently noted by experiments conducted at BBN[7]. The reported approach to dealing with this problem is to renormalize the warp distribution such that the histogram median is shifted towards 1.0 at each iteration. If the median warp after an iteration is for example 1.04, then all warp factors are adjusted by substracting 0.04 before starting the next iteration. Given the comparable performance gains reported by BBN, this seems a reasonable approach to solve this drifting problem.

The performance of speaker normalization could be improved by making the warp factor phone dependent rather than speaker dependent. It is questionable however if the current likelihood based approach is suitable as is in such a framework as the current approach requires much more data than the average duration of a phone to make a reliable estimate of the warping factor as illustrated in figure 35 for a few speaker. As shown, most speakers required approximately 10 seconds of data before a "converged" warp factor estimate was found.

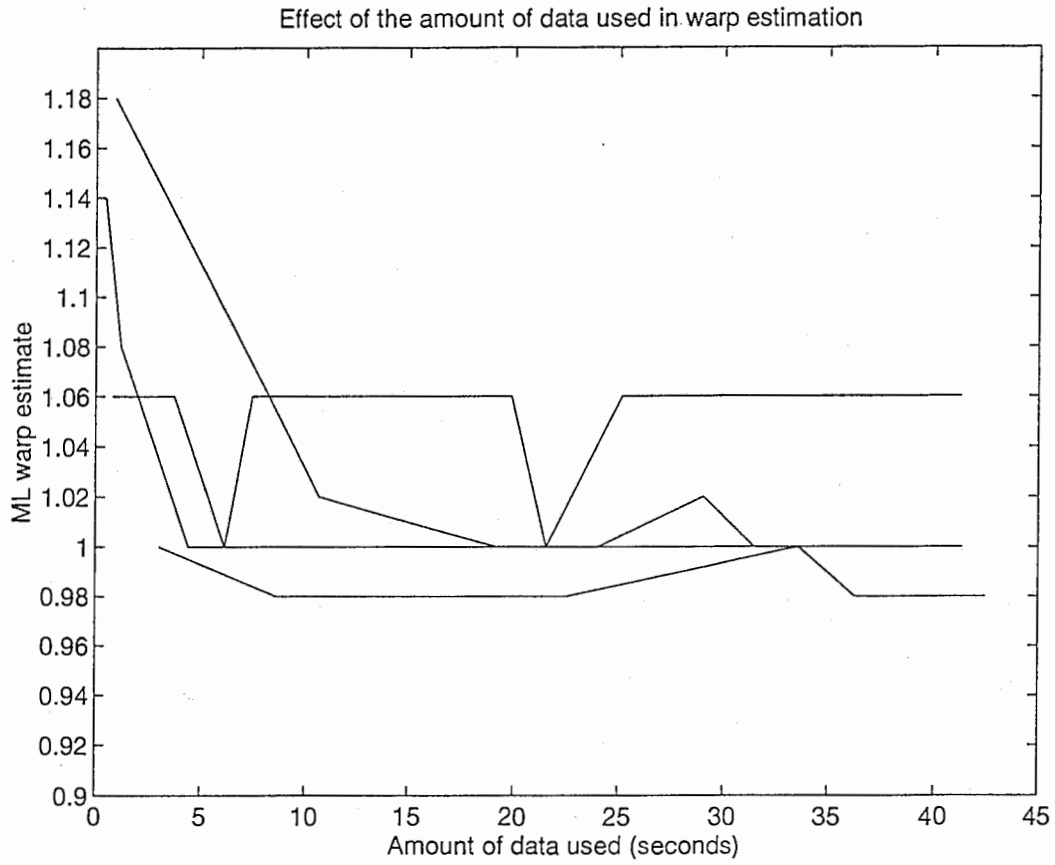

Figure 35: ML estimate of the warping factor for 4 different speakers as a function of the amount of data used in estimation

References

[1] E. Eide and H. Gish, "A Parametric Approach to Vocal Tract Length Normalization", *Proc. of the Int. Conf. on Acoust., Speech and Signal Proc.*, pp. 346-348

[2] P. Zhan and M. Westphal, "Speaker Normalization Based on Frequency Warping", *Proc. of the Int. Conf. on Acoust., Speech and Signal Proc.*, , pp. 1039-1042, 1997

[3] T. Kamm, G. Andreou and J. Cohen, "Vocal Tract Normalization in Speech Recognition: Compensating for Systematic Speaker Variability", *Proc. of the 15th Annual Speech Research Symposium*, pp. 161-167, CLSP, Johns Hopkins University, Baltimore, MD, June 1995

[4] R. Roth, L. Gillick, J. Orloff, F. Scattone, G. Gao, S. Wegmann and J. Baker, "Dragon systems' 1994 Large Vocabulary Continuous Speech Recognizer", *Proc. ARPA Workshop on Spoken Language Technology,* pp. 116-210, Austin, TX, January 1995

[5] L. Lee and R. Rose, "Speaker Normalization using Efficient Frequency Warping Procedures", *Proc. of the Int. Conf. on Acoust., Speech and Signal Proc.*, , pp. 353-356, 1996

[6] S. Wegmann, D McAllaster, J. Orloff and B. Peskin, "Speaker Normalization on Conversational Telephone Speech", *Proc. of the Int. Conf. on Acoust., Speech and Signal Proc.*, , pp. 339-341, 1996

[7] Miller *et al.*, BBN submission, LVCSR meeting, April 1997, Baltimore, MD

[8] M. Bacchiani, "The ASSM Toolkit for Polynomial Segment Models and Automatic Unit Design", ATR technical report TR-IT-0225, June 1997

[9] M. Ostendorf, M. Bacchiani, Y. Sagisaka and K. Paliwal, "Speech Recognition System Design using Automatically Learned Non-Uniform Segmental Units", ATR technical report TR-IT-0147, January 1996