MASTER THESIS

REGION-BASED VOCAL TRACT LENGTH NORMALIZATION FOR AUTOMATIC SPEECH RECOGNITION

by

MICHAIL MARAGAKIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE AT TECHNICAL UNIVERSITY OF CRETE

Crete, Creece 2008

The undersigned certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled

"REGION-BASED VOCAL TRACT LENGTH NORMALIZATION FOR AUTOMATIC SPEECH RECOGNITION"

by Michail G. Maragakis

in partial fulfillment of the requirements for the degree of Master of Science.

Michail Maragakis: *Master Thesis,* REGION-BASED VOCAL TRACT LENGTH NORMALIZATION FOR AUTOMATIC SPEECH RECOGNITION, 2008

COMMITTEE OF EXAMINERS:

Alexandros Potamianos

Vasilios Digalakis

Athanasios Liavas

To Tzimi P.
(Sorry for the delay Tzimi)

ABSTRACT

In this work, normalization techniques in the spectral domain which result to the improveness of automatic speech recognition systems are studied.

Normalization is basic strategy in order to reduce the mismatch. Common normalization schemes proposed in the literature are motivated and discussed, and promising method is implemented and studied in detail.

Vocal tract length normalization is a popular technique for (unsupervised) speaker normalization especially when small amounts of speaker data is available. Frequency warping approaches to vocal tract length normalization have been proposed and evaluated on speech recognition tasks. These approaches are based on warping the frequency axis, parameterized by a scalar warping factor and a single warping function. This thesis presents a frequency warping method that is based on dividing unsupervisedly test utterance' s frames into regions and warping independently each region's frames. The proposed method is presented in the context of two-pass existing methods for frequency warping based speaker normalization. Performance improvements obtained using the newly proposed method are shown to increase word accuracy when applied to subsets of AURORA4 Speech Corpora under clean conditions.

We have experimented with standard mono-parametric linear warping VTLN algorithms. Additionally, we investigate alternative warping functions, phone-dependent warping functions, as well as combinations of warping and maximum likelihood bias removal. For this purpose, we investigate warping functions that minimize the spectral distance between two speaker's utterances.

The study of the phonemes's behaviour during warping is the first task that will be studied in this thesis. For these initial experiments, we use the TIMIT database. VTLN maps from the reference to the mapped speakers. First the effectiveness (in terms of MSE reduction) of linear, power and piecewise-nonlinear frequency warping function is investigated. Next, bi- and four-parametric warping functions are investigated; both phone-independent and phone-dependent warping algorithms are evaluated.

After the study of the dependence between warping and the various phonemes is investigated and based on the extracted conclusions, a mechanism for the division of test utterance's frames in regions and the estimation of an optimal, for each region, warping factor and function is provided. The standard two-pass recognition method is extended so that region-dependent optimal warping factor and function can be obtained from a set of candidate factors and functions, based on ML criterion and through a grid search over these two sets.

It will be shown that this formalism is an extension of the unique level formalism of factor estimation. The new added levels are determined by the number of the

candidate warping functions and regions. Secondly, it will be shown that for cases that the number of regions is growing, constraints that are taken into account are consistent with the results extracted from cases that the number of regions is few and critical for combining good performance and computational efficiency.

ACKNOWLEDGMENTS

At this point, I would like to express my thanks to the people who supported and accompanied me during the progress of this work.

I would like to thank my supervisor, Associate Prof. Alexandros Potamianos for giving me a deep insight into automatic speech recognition. He gave me the opportunity to pursue my ideas, he followed my work with continuous interest and he supported me with numerous ideas and discussions.

I am indebted to all my colleagues at the Telecommunication Group for their discussions, comments and hints.

M. M.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

VTLN *V*ocal *T*ract *L*ength *N*ormalization

MFCC *M*el *F*requency *C*epstral *C*oefficients

HMM *H*idden *M*arkov *M*odel

VTL *V*ocal *T*ract *L*ength

DCT *D*iscrete *C*osine *T*ansformation

ASR *A*utomatic *S*peech *R*ecognition

MSE *M*ean *S*quare *E*rror

1 INTRODUCTION

Automatic Speech Recognition (ASR) provides a means of communication between humans and machines. ASR systems try to achieve human-like performance in recognition. However, human-like performance is still a target because of several variability reasons that the speakers provide. The current proposed algorithms remains unable to handle various kinds of unwanted variabilities observed in the speech signal. These variabilities in speech arise due to several factors including the differences in the speech production mechanisms of the speakers such as Vocal Tract Length (VTL). This leads to a non-robust recognition performance of the ASR systems when exposed to different conditions.

It is generally known that one of the major source of inter-speaker variability is the vocal tract shape, more specifically the VTL. The length of the vocal tract can vary from approximately 13 cm for adult females to over 18 cm for adult males. This source of variability results in a significant degradation in recognition performance. It has been found that vocal tract length variation causes scaling in the spectral domain since the formant frequencies are inversely proportional to the length of the tube. Many normalization schemes try to eliminate the variability by re-scaling the frequency axis resulting in substantial improvementes in speech recognition performance.

Vocal Tract Length Normalization (VTLN) tries to compensate for the effect of speaker-dependent vocal tract lengths by warping the frequency axis. This thesis aims to improve the robustness of the ASR systems by handling the effects of the vocal tract length. Differences in vocal tract conditions could arise among speakers. The current thesis tries to propose methods normalizing this variability. The contribution of this thesis proposal is to describe a set of frequency warping based speaker normalization techniques. As a first step, we have studied the relationship between phonemes and warping methods in order to understand the nature of warping influence between phonemes. This dependence can lead us to a better phoneme categorization in order to remove individual speaker charactiristics and, thus, improve the recognition performance. Also the effects of bias addition to the unwarped specturms are analyzed. After that, based on the extracted results, a locally contrained VTLN method is proposed. This technique which can improve the robustness of the ASR system against the variability factors has been proposed, investigated and evaluated in this thesis.

1.1 THESIS ORGANIZATION

This thesis is organised as follows. First, the robust speech recongition systems are described in Chapter 2. After that, in Chapter 3, existing frequency warping based speaker normalization procedures are described. In Chapter 9, there is a decription of the training and testing subsets of databases that we used in this thesis. In Chapter 4, we examine the ability of various warping functions to reduce the spectral distance between speakers for various phonemes (vowels and others). In Chapter 9, there is a decription of the training and testing subsets of databases that we used in this thesis. Chapter 5 describes the proposed region-based VTLN algorithm that first categorizes the testing utterance's frames into regions and then region-specific spectral warping functions and factors are computed using an ML criterion in order to optimally warp each region's frames. Chapter 6 presents the results of the experimental study, comparing the Region-Based VTLN with already existing normalization techniques. Finally, discussion and summary are provided in Chapter 7.

2 ROBUST AUTOMATIC SPEECH RECOGNITION

Approaches for ASR are based on statistical representations of the speech signals. ASR involves :

- Feature extraction and
- Statistical modeling at extracted, from feature extraction, vectors.

Feature extraction computes a sequence of vectors representing the linguistic information in the speech signal. Feature extraction discards unwanted variabilities by transforming the signal to spectral and, in order to reduce the dimensionality of the spectral vectors, to cepstral coefficients.

Statistical modeling estimates likelihood of match between the extracted vector sequence and a set of reference probability density functions in order to facilitate sentence decoding. The strategy which is followed is that the word sequences are divided into smaller segments, with the total number of distinct segments being restricted to a finite number. Typically such a segmentation is done at the phonetic level.

The feature extraction and statistical modeling of the ASR systems are explained with more details in the following subsections.

2.1 FEATURE EXTRACTION

The speech signals generated by humans are continuous-time signals. For the processing of these signals by the machines, which can do only a digital processing, the signals are digitized by an analog to digital (AD) converter. AD converter outputs the digital version of continuous-time signal by sampling and then quantizing the amplitudes. Telephone speech is the most common speech used in ASR systems, whose bandwidth is typically from 200 Hz to 3400 Hz. According to Nyquist sampling theorem, minimum sampling frequency for AD conversion of a signal should at least be twice the maximum bandwidth of the signal, to avoid aliasing of the signal (an effect that avoids the perfect reconstruction of the continuous-time signals from the digitized signal). Hence, the typical sampling frequency used for sampling of the speech signals is 8000 Hz.

Signal analysis is an autonomous part of modern speech recognition architectures. It is based on short-term spectral analysis (basic principles of spectral analysis are described in [16]).

Throughout this work, the MFCC-based HTK signal analysis front-end will be used [2]. First, it is common practice to pre-emphasize the speech signal by

applying the first order difference equation to each of the samples of each window. The pre-emphasis coefficient k is equal to 0.97. After preemphasis, it is usual to taper the samples in each window so that discontinuities at the window edges are attenuated. This is done by using the Hamming window. Most of the features used for speech recognition are based on Fourier analysis of the signals. Fourier analysis requires the characteristics of the signal taken for analysis to be stationary throughout. But speech signals are, in general, nonstationary. However, from the knowledge about the human speech production system, inertia of the articulators do not allow the characteristics of the speech signal to change rapidly over time. In other words, the characteristics of the signal can be approximated to be stationary over a short period of time segments. Hence, for further processing, the speech signal is divided into a sequence of short signals called frames, by performing a sequence of shifting and windowing operation on the original signal.

The preemphasized speech signal is segmented into windows of 25 ms length using the Hamming window. These windows are overlapping each other for 15 msec. That means that we finally take frames length 10 msec. The underlying idea is that the speech signal is quasi stationary for 20 to 50 msec, which supports short-term spectral analysis within a window of 25 msec length.

Typical features used for speech recognition are based on the power spectral representation of the speech signal [3]. The power spectrum is computed by finding out magnitudes of the complex-valued Fourier coefficients obtained from the Discrete Fourier Transform (DFT) of the speech frames. If N represents the frame length and $s = s[0], s[1], ..., s[N-1]$ represents the speech signal, then the DFT coefficients $S[k]$ can be computed by equation

$$S[k] = \sum_{n=0}^{N-1} s[n] exp\left(j\frac{2\pi}{N}kn\right), 0 \leq k \leq N-1$$

The magnitude coefficients are then binned by correlating them with each triangular filter. Here binning means that each FFT magnitude coefficient is multiplied by the corresponding filter gain and the results accumulated. The human ear resolves frequencies non-linearly across the audio spectrum and empirical evidence suggests that designing a front-end to operate in a similar non-linear manner improves recognition performance. A popular alternative to linear prediction based analysis is therefore filterbank analysis since this provides a much more straightforward route to obtaining the desired non-linear frequency resolution. The frequency axis is warped according to the Mel-scale [30]. The filters are triangular and they are equally spaced along the mel-scale which is defined by,

$$Mel(f) = 2595 \cdot log_{10}(1 + \frac{f}{700})$$

As a result of this, the spectral resolution is reduced towards higher frequencies similar to the frequency response of the human ear.

Usually some external knowledge about the human perception system or human speech production system is utilized to transform the power spectrum to feature vectors. During such transformation, the main aim is to emphasis the linguistic information and suppress the unwanted variabilities present in the power spectrum. Features extracted from power spectrum, that are shown to be successful for ASR are the mel-frequency cepstral coefficients (MFCC) [8]. For the evaluation of MFCC coefficients, the power spectral values are integrated within overlapping mel-scaled critical band windows to obtain what is called mel-scaled critical bank spectrum. The critical band spectral amplitudes are then compressed by a logarithmic function. The resultant values are then transformed through an inverse discrete cosine transformation (DCT) to obtain the MFCC coefficients. The highest cepstral coefficients are omitted because they contain only little information about the spoken word sequence. The resulting vector of typically 12 coefficients are the standard MFCC vectors. Speech recognition systems often incorporate the temporal dynamics of the speech signal in the feature representation by including the first and second derivatives of the static feature vectors. For the case of MFCC vectors, these are augmented with the first derivatives of all cepstral coefficients and the second derivative of the zeroth or energy cepstrum coefficients. These augmented coefficients are computed by linear regression from three and five successive cepstrum vectors. The size of the final acoustic vector is the desired dimension of typically 39 coefficients.

2.2 ACOUSTIC MODELLING.

After the feature extraction of training vectors, we train the acoustic models. Denote that $Y = y_0, y_1, ..., Y_N$ represents the set of feature vectors extracted from the speech signal, statistical modeling techniques formulate the speech recognition problem as a maximum a posteriori (MAP) problem. The aim is to provide the probability that each hypothesized word sequence W generates given an observed sequence of acoustic vectors Y.

$$W^* = argmax_W P(W/Y) \tag{2.1}$$

The W^* is the most likely word sequence from the set of all possible word sequences, given X. Also it is the recognition result. The MAP formulation of speech recognition is hard to deal with directly. It is usually reformulated into a

problem based on likelihood estimation using Bayes rule, as follows:

$$\begin{aligned} W^* &= argmax_W P(W/Y) \\ &= argmax_W \frac{P(Y/W)P(W)}{P(Y)} \\ &= argmax_W \frac{P(Y/W)P(W)}{P(Y)} \\ &= argmax_W P(Y/W)P(W) \end{aligned} \tag{2.2}$$

The denominator $P(Y)$ is independent to candidate word sequences and so it is omitted taken account that we want the argument W which maximizes the probability. In the above equation, $P(Y/W)$ denotes the acoustic model and the $P(W)$ denotes the language model.

Hidden Markov models are stochastic finite state automata. They consist of a number of states and transitions between these. Each state is characterized by the probability to observe a given acoustic vector (emission probability), and the probability to step into one of the possible successor states (transition probability). Assuming the feature vector sequence Y is a stationary process that has been generated by a sequence of HMM states, denoted by $X = x_1, x_2, ..., x_N$, an acoustic model θ is the sum of all hidden Markov model parameters that describe the sub-word units of a speech recognition system, $p(Y/W) = p(Y/W;\theta)$.

To make the model simple and computationally tractable, simplifying assumptions are made while applying HMMs to the acoustic modeling problem. More specifically, the usual Hidden Markov Model which is used is the First Order Hidden Markov Model. In Figure , we may see the topology of this model.

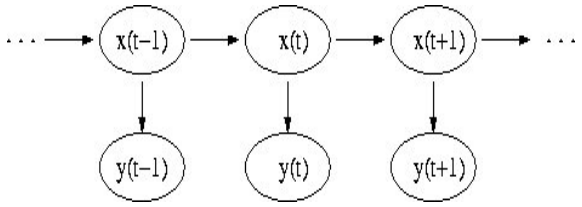


Figure 1. Topology of First Order Hidden Markov Model.

Based on this topology and for every moment t, we extract the following assumptions :

- $x(t+1)$ independent to $x(t-1)$ given to $x(t)$
- $y(t)$ independent to $x(t-1)$ given to $x(t)$
- $y(t)$ independent to $y(t-1)$ given to $x(t)$,$x(t-1)$

Under these assumptions, the acoustic model probability becomes:

$$P(Y/W) = \sum p(x_0) \prod_{t=1}^{T-1} p(x_t/x_{t-1}) p(y_t/x_t) \tag{2.3}$$

where x_0 denotes the initial state.

Most common for emission density modeling is to use Gaussian Mixture Model (GMM). GMM is a weighted mixture of several Gaussians. It is characterized by the weighting factors, mean vectors, and covariance matrices of all the Gaussians. The expression for density function for GMM is given by,

$$p(y) = \sum_{k=1}^{K} c_k G_k(y)$$

where K denotes the number of Gaussians in the GMM, and c_k denotes the weighting factor for k'th Gaussian, G_k. If μ_k and Σ_k denote respectively the mean vector and covariance matrix of the k'th Gaussian, and if D denote the feature vector dimension, the expression for $G_k(Y)$ is given by,

$$G_K(Y) = \frac{1}{2\pi^{\frac{D}{2}} |\Sigma_k|^{\frac{1}{2}}} e^{-\frac{1}{2}(y-\mu_k)^T \Sigma_k^{-1} (y-\mu_k)}$$

2.3 LANGUAGE MODELING

A language model gives the probability $P(s)$ of a sentence s. Let S be a word sequence and M be some underlying structure related with it. Statistical language modeling estimates $P(S)$, while computational linguistics deals with the estimation of $P(S \mid M)$ [35].

The majority of the language models decomposes the sentence probability, $P(s)$, into a product of conditional probabilities

$$P(s) = P(w_1 \dots w_n) = \prod_{i=1}^{N} P(w_i | h_i) \tag{2.4}$$

where w_i is the i^{th} word in the sentence and $h_i = \{w_1, w_2, \dots, w_{i-1}\}$ is the sequence of preceding words.

2.4 N-GRAM LANGUAGE MODELING

The simplest language probabilistic model let any word to follow any other word with equal probability. For example, if the vocabulary of a certain natural language

consists of 75.000 unique words, then the probability of any word following any other word equals to $\frac{1}{75.000}$. A more complex language model uses the frequency of occurrence of a word. For example, the previous paragraph has totally 81 words, in which the words "word" and "a" occur 3 and 5 times, respectively. According to the simple language model the words "word" and "a" have $\frac{3}{81}$ and $\frac{5}{81}$ probability, respectively, to follow any word. But for the sequence "to predict the next", the word "word" is more reasonable than "a" to follow "next". This intuitive observation considers the conditional probability of a word given the previous word, instead of using the relative word frequency.

The N-gram language model considers the language as a Markov process of order $N-1$.

$$P(w_i|h_i) = P(w_i \mid w_{i-N+1}, \ldots, w_{i-1}) \approx P(w_i \mid w_{i-N+1}^{i-1}) \tag{2.5}$$

Equation 2.5 states that the probability of word w_i given all the previous words of the sentence can be approximated by the probability given only the previous $N-1$ words.

N-gram probabilities are computed by counting and normalizing the N-gram occurrences. For the bigram case the conditional probability of word w_{i-1} given that it is followed by word w_i is computed as

$$P(w_i|w_{i-1}) = \frac{C(w_{i-1}w_i)}{\sum_w C(w_{i-1}w)} = \frac{C(w_{i-1}w_i)}{C(w_{i-1})} \tag{2.6}$$

Equation 2.6 takes the count of $w_{i-1}w_i$ bigram and divides it by the sum of all bigrams that have w_{i-1} as first word. Note that the latter sum is equal to the count of w_{i-1} unigram. For the general case of N-gram model the above equation is written as

$$P(w_i|w_{i-N+1}^{i-1}) = \frac{C(w_{i-N+1}^{i-1}w_i)}{C(w_{i-N+1}^{i-1})} \tag{2.7}$$

Equations 2.6 and 2.7 use the frequency interpretation of probability [36], applying the technique of Maximum Likelihood Estimation (*MLE*). Even with large corpora many N-grams occur only once or they have low counts, so, the computation of N-gram probabilities remains a sparse estimation problem. This fact is prescient with Chomsky's observation that a model suffering from lack of data assigns low probability to a phrase regardless its sensical or grammatical correctness [33]. Thus, it is preferable not to apply *MLE* of N-gram probabilities in a straightforward way, based on counts. Instead, several smoothing approaches [34] can be used in order to smooth the *ML* estimates.

2.5 SPEECH VARIABILITY

Differences in VTL can partially explain why some people have deeper voices than others. Knowledge of the gender is important here as adult male speakers

have longer VTL than adult female speakers. The differences in VTL create a clear division between male and female speakers. The location of these resonance frequencies depends directly on the length of the tube. Variations of the tube's length will shift the resonance frequencies along the frequency axis. For instance, the longer the tube gets, the more resonance frequencies are shifted to the lower end of the spectrum. Decrease the length and the frequencies will be shifted to higher frequencies.

The inter-speaker differences in speech are partly due to differences in speakers anatomy especially in the Vocal Tract geometry. More precisely, the VTL creates variations in the resonance frequencies of identical phonemes. Spectral analysis will clearly reveal the resonance frequencies's location on the frequency axis.

2.5.1 *Intra-Speaker Variability*

Intra-Speaker Variability [4] stems from the natural randomness in the pronunciation of the smallest constituents of speech, the phonemes [9, 12, 11, 18, 20]. A person will rarely produce the same phoneme in an identical manner twice. Furthermore, speech is constituted of a series of phonemes, each one of them being pronounced differently depending on the neighboring phonemes. This is known as the coarticulation effect. Further, the intra-speaker variability combines all variations of speech, including the effects of mood, stress or even health.

2.5.2 *Inter-Speaker Variability*

Inter-Speaker Variability accounts for the fact that speech is different among speakers [4]. For instance two individuals will not produce the same speech even if they are asked to say the same sentences. Differences in size, age, gender, speaking rate and accentuation are sources of variations on the phoneme level between individuals. Regional and local accents are also sources of variability among the same language.

3

VOCAL TRACT LENGTH NORMALIZATION

Vocal Tract Length Normalization (VTLN) is a speaker normalization algorithm which improves automatic speech recognition (ASR) performance. The algorithm compensates for the effect of speaker's vocal tract length by warping the frequency axis of the spectrum magnitude before computing the coefficients at cepstral domain [12].

Most papers published subsequently about vocal tract length normalization addressed one or more of the following topics:

- type of the frequency axis warping function (linear, non-linear) and its implementation (time domain, frequency domain, cepstral domain)
- reliable estimation of the warping factors in training
- efficient warping factor estimation in test (with respect to word error rate, required adaptation data, and computational overhead)
- gain in recognition accuracy achieved by VTLN under different conditions (clean vs. noisy environment, small vs. large training corpora, small vs. large vocabulary)
- comparison of VTLN with adaptation techniques, sequential application of VTLN and adaption schemes (e.g. MLLR)

VTLN is applied using warping functions that depend only on a few free parameters [15]. Even with one parameter, the warping factor α and using typically a single utterance, VTLN performs well for a variety of recognition tasks. This unique parameter can be evaluated by calculating formants frequencies [10] or by using a maximum likelihood (ML) criterion usually in a two-pass speech recognition scenario [12, 21, 22].

Li Lee and Richard Rose proposed [12] a set of low complexity, maximum likelihood based frequency warping approach to speaker normalization. Lee et al. [12] proposed an efficient maximum likelihood algorithm for estimating the warping factor for linear frequency scaling. They estimated warping factors in a maximum likelihood framework. For training speakers, an iterative procedure was proposed, whereby an acoustic model was trained on one half of the normalized training data, which was then used to estimate warping factors for the other half. Subsequently the data sets were swapped and the warping factors for the first half were re-estimated with a new acoustic model trained on the second half of data. It was found that more than one iteration reduced the word error rate on the training data, but not anymore on the test data.

According to [12], for each utterance an optimal warping factor $\hat{\alpha}$ is selected from a discrete ensemble of M possible values so that the likelihood of the warped utterance is maximized with respect to a given speech recognition model (ensemble of hidden Markov models) and a given transcription. The transcription is obtained from a first recognition pass. Instead of re-sampling the speech waveform in the time domain, Lee and Rose proposed furthermore to incorporate linear frequency axis warping

$$\omega \rightarrow \tilde{\omega} = g_{\alpha}(\omega) \tag{3.1}$$

into Mel-frequency warping by modifying the center frequencies and bandwidths of the filter bank channels. The ω and $\tilde{\omega}$ are correspondingly the unwarped and warped frequencies and g is the linear warping function.

The optimal warping factor is obtained by searching over a grid of 13 factors spaced evenly between $0.88 \leq \alpha \leq 1.12$ with spaces 0.02 between the factors.

Let $X^{\alpha} = g_{\alpha}(X)$ denote the sequence of cepstral vectors where each one of them is warped by the warping function $g_{\alpha}()$. If λ denotes the parameters of the unnormalized HMM models and W is the transcription obtained from an initial recognition pass, the optimal warping factor (referred as *global* henceforth) is defined as

$$\hat{\alpha}_{glb} = \arg\max_{\alpha} P(X^{\alpha}|\lambda, W) \tag{3.2}$$

After the estimation of the $\hat{\alpha}_{glb}$, the frequency warped observation vector $X^{\hat{\alpha}_{glb}}$ is decoded in a second recognition pass to obtain the decoded transcription.

On a telephone based connected digit recognition task, Lee and Rose achieved a reduction of word error rate of 15% relative.

Alexandros Potamianos and Richard Rose [1] demonstrate that frequency warping combined wtih ML speaker adaptation can gain advantage on the performance. The simultaneous warping at the frequency domain and reshaping the spectral energy contour can reduce the error rate in a telephone based connected digit task by 30-40%.

In other methods, VTLN was used during both the training and the testing procedure [21]. In 1996, Eide and Gish [9] investigated the impact of different frequency warping functions and of the amount of training data on the recognition performance with VTLN. Based on the median position of each speaker's third formant, the corresponding factors were estimated. The differences to linear and nonlinear warping were small, with the non-linear warping function reports better performance. Concerning the amount of training data, on the SwitchBoard corpus, a reduction in word error rate (WER) of 8% relative was obtain when 5 hours of training data were used, which reduced to 6% relative when the full training corpus of 63 hours was utilized. Eide et al. [9] investigated also methods to enrich the training corpus with additional normalized data, but they could not reduce the word error rate any further.

Wegmann et al. [18] applied a piece-wise linear warping function and proposed a fast warping factor scheme by training one model of normalized speech. This model was trained in an iterative mode. The obtained speech model was used to estimate the warping factors. The normalized data were used for retraining a new model. Only voiced frames were used for warping factor determination. Evaluating on the SwitchBoard corpus, the WER were reduced by 12% relative for gender-independent and by 6% relative for gender-dependent acoustic modeling.

Zhan and Westphal [23] based on the median position of the first three formants with maximum likelihood estimates, compared warping factor estimation. They concluded that this approach consistently outperformed other estimates which were based also on formants. A piecewise-linear warping function yielded better results than the non-linear function proposed by Eide and Gish [9]. In order to reduce the computational cost of grid search over all warping factors in two-pass VTLN, Zhan and Westphal proposed to keep the alignment between acoustic vectors and HMM states from the first recognition pass fixed when performing the grid search for the best warping factor. In the optimal setup, the word error rate could be reduced by 9% relative on a 5k-word vocabulary Spanish spontaneous speech scheduling task.

Gouvea and Stern [10] proposed a warping factor estimation based on the median frequency of the first three formants. They fitted a linear warping function. In return they got consistently better results in clean and noisy conditions with word error rate reduction of up to 15% on the Resource Management database. Five sentences at minimum were required to estimate reliably the warping factor.

Welling et al.[27] investigated VTLN and MLLR on the Wall Street Journal corpus with a 5k-word vocabulary test set. Based on two-pass recognition, the word error rate was reduced by 11% relative in gender-independent, and by 4% relative in gender-dependent mode. On another database, the German SieTill, consisting of connected digit strings, the gain of VTLN increased when simple acoustic models were used. Finally, they proposed an alternative scheme for fast warping factor estimation in test based on one Gaussian mixture model for normalized speech similar to the technique proposed by Wegmann et al. On the WSJ corpus, this approach performed almost as good as two-pass VTLN.

Welling et al published further results with fast warping factor estimation [21]. Using a simple method, they omitted silence frames from the warping factor estimation based on the observation counts of each density in the Gaussian mixture model. In addition, they suggested a simplified non-iterative maximum likelihood method for warping factor estimation in training. They found that a single densities acoustic model gave better results than more complex mixture density models. The two-pass approach could be improved by using unnormalized acoustic models for the first recognition pass, which on the other hand increased the gap between two-pass and fast VTLN. A word error rate reduction of 9% relative was achieved with Gaussian mixture model based fast VTLN on the 5k-word vocabulary Wall Street Journal test set, whereas two-pass VTLN yielded

up to 17% relative WER reduction. On the German spontaneous speech task VerbMobil I, the reduction was 5% relative at best.

Westphal et al. [24] compared maximum likelihood warping factor estimation with a new criterion based on linear discriminant analysis. The new criterion lead to a faster convergence in iterative warping factor estimation of training data, and the derived speaker cluster were more discriminant. With respect to the word error rate, the new criterion performed slightly worse on the German VerbMobil I task and slightly better on a Chinese dictation task.

Dolfing [25] evaluated the efficiency of two maximum likelihood criteria for warping factor estimation. One was text-dependent similar to two-pass VTLN with the preliminary transcription replaced by the reference transcription in a supervised manner. The other resembled the GMM based text-independent warping factor estimation of Lee and Rose. Based on an internal dictation database he compared the word error rate obtained by these techniques with an optimal error rate by choosing the warping factor with the lowest word error rate. If warping factors were estimated in a speaker-wise fashion, the text-dependent criterion yielded about 9% of the maximum possible reduction in WER. Sentence-wise estimation of the warping factor left more room for improvements, which is why is that case only about 7% of the maximum possible gain was achieved. Preliminary experiments with the text-independent estimation indicated that its performance was only slightly inferior to the text-dependent technique, but conclusive recognition result were not reported.

Cox [28] presented a method to implement VTLN at the cepstrum stage. As Mel-frequency axis warping is approximately a logarithmic scaling of the frequency axis, linear frequency axis warping amounts to a constant frequency shift in the Mel-frequency domain. This fact was used to derive a transformation matrix that compensates for the shift in the cepstrum domain. The functional form of this type of frequency axis warping was similar to highly constrained MLLR with only four free parameters. Phoneme recognition tests in supervised normalization mode using the Wall Street Journal database showed reduced error rates only if the means of single-density acoustic models were adapted. A normalization of the test data did not yield the same improvement. A minor additional gain was found when a different amount of warping was allowed at different spectral bands. The overall reduction in phoneme error rate was 4% relative at best.

In 2001, Pitz et al. showed that VTLN equals a linear transformation in the cepstrum domain for arbitrary invertible frequency axis warping functions [6]. Yet another approach for fast warping factor estimation was presented in the same year by Emori and Shinoda [7]. They applied bi-linear frequency axis warping in the cepstrum domain and proposed an approximation to compute the warping factor from cepstral coefficients. On a Japanese isolated-word recognition task they achieved similar performance in supervised mode like maximum likelihood estimation at smaller computational costs. Better results were achieved if only vowels were used for estimation, but comparable result for maximum likelihood

estimation were not given.

Analyzing the speech, it becomes known that spectral differences among speakers due to vocal tract length are both phone-dependent and non-linear and cannot be fully captured by a single warping function and factor selected on a per utterance basis. Recently, there have been attempts to compute "instantaneous" warping factors, i.e., warping factors on a per frame basis. Among these frame-based VTLN approaches the most notable are the MATE framework [13] where spectral warping is applied to individual frames using a two-dimensional Viterbi decoding algorithm to estimate the frame-based warping factor. This efficient and effective method is used for compensating for local variability in speech which may have potential application to a broader array of speech transformations. The techniques are presented in the context of existing methods for frequency warping based speaker normalization and existing methods for computation of dynamic features for ASR. The modified decoding algorithms were evaluated in both clean and noisy task domains using subsets of the Aurora 2 and Aurora 3 Speech Corpora under clean and noisy conditions. It was found that, under clean conditions on the Spanish Language Subset of the Speech-Dat- Car database, the modified decoding method applied with local frequency transformations reduced word error rate (WER) by 24 percent. This was a factor of two greater reduction in WER than was obtained on the same task using the more well known frequency warping based vocal tract length normalization (VTLN) procedure. Furthermore, the MATE decoder can be applied to select the frame specific temporal resolution for the dynamic features in MFCC feature analysis. The computation of dynamic features through linear combination of successive MFCC feature vectors is important for modeling the time evolution of spectral information in speech. Allowing local optimization of the temporal resolution over which the first and second order difference cepstral are computed is equivalent to a non-uniform sampling of the time scale for dynamic features.Temporal resolution for the dynamic features to be estimated as part of search.

Shin et al [19] selected the best warping factor based on a normalized codebook. Shin et al presented is an iterative method of constructing the "normalized" codebook that can be used as a text independent warp factor estimator for LVCSR system. Given the normalized codebook, the warp factor is estimated by searching the best fitting warped version of feature vectors of a given utterance. Throughout the whole process of normalized codebook construction and warp factor estimation, neither acoustic, nor phonetic knowledge is made use of. The effectiveness of the proposed method is investigated by performing recognition experiments. Given an initial codebook trained with unwarped feature vectors, the normalized codebook is derived by means of a hierarchical two level iterative refinement processes: Progressive Refinement Process (PRP) and Local Minimization Process (LMP). By exploiting these processes, the resulting codebook has the effect of having been trained with the normalized training vectors even though no reference of normalized vocal tract length is explicitly used. The results showed more than

4% improvements in word level accuracy.

Sankaran Panchapagesan [15] developed a gradient search algorithm for VTLN estimation with MFCC features. The novel calculation was that of the gradient gradient of the filterbank with respect to the FW parameters. For male children speakers tested on models trained from adult males, the algorithm was used to estimate multiple-parameter FW for VTLN and more than 50% relative reduction in word error rate was obtained compared to single-parameter PL VTLN. For single parameter PL VTLN, the algorithm was more efficient than the widely used grid search by a factor of around 1.6. For multiple parameters, grid search would be inefficient and the computational savings of gradient search would be greater.

Jonas Loof, Hermann Ney and Srinivasan Umesh [31] presented an flexible approach to VTLN warping factor estimation. Due to the equivalence of frequency warping and linear transformation of cepstral coefficients, warping factors can be efficiently estimated by accumulating the sufficient statistics for linear transformation estimation and searching the constrained space of transformations given by the explicit mapping between warping factors and linear transformation matrices.

S. V. Bharath Kumar et al, [32] using formant data from Peterson & Barney and Hillenbrand vowel databases, analyzed the behavior of scale factor as a function of frequency. The empirical observation showed that while uniform scaling assumption may be valid at higher frequencies, there were significant deviations at low frequencies. They showed that while our recently proposed model had behavior similar to the empirical result, the behavior of many of the commonly used non-linear models differ significantly from the empirical result. They also showed that our proposed model does better fitting to the formant data than these non-linear models. They concluded that the affinetransformation model may be a more appropriate non-linear model for speaker normalization.

4

DEPENDENCE BETWEEN PHONEMES AND WARPING

Based on the fact that VTLN is a popular technique for unsupervised speaker normalization especially when small amounts of speaker data are available, we investigate alternative warping functions, phone-dependent warping functions, as well as combinations of warping and maximum likelihood bias removal. At this chapter, we investigate warping functions that minimize the spectral distance between two speakers's utterances. The proposed method uses the Mean Square Error (MSE) metric and linear, piecewise linear and nonlinear warping functions. VTLN in combination with linear bias removal is also investigated.

For our initial experiments, we use the TIMIT database. More specifically, we select 16 speakers, 8 male and 8 female, from the TIMIT training set. The speakers are separated into "reference" (8 speakers, 4 male and 4 female) and "mapped" speakers (the rest 8 speakers). VTLN maps from the reference to the mapped speakers. First the effectiveness, in terms of MSE reduction, of linear, power and piecewise-nonlinear frequency warping function is investigated. Next, mono-, bi- and four-parametric warping functions are investigated. Both phone-independent and phone-dependent warping algorithms are evaluated.

4.1 WARPING INFLUENCE.

In this section we will investigate the influence that warping introduces to the phonemes (vowels and others). Potamianos et al. [26] have implementing linear frequency warping at the phoneme level and cepstral domain. Introducing as distance criterion the MSE distance, the similarity between two frames before and after warping is studied. Given the unwarped spectrum X (reference spectrum), the warped spectrum Y (mapped spectrum) and the warping function g_α, the MSE is defined as,

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\Big(X_i - g_\alpha(Y_i)\Big)^2$$

where N=256 is the number of mel coefficients at spectrum domain. The frequency warping is performed as follows:

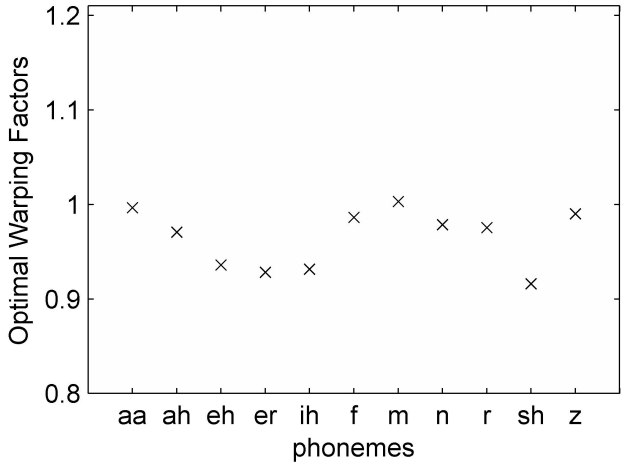
- For each phoneme and speaker and for the middle frame of the utterance, the average spectral envelope is computed,
- An optimal warping factor $\hat{\alpha}$ is computed so that the MSE between the warped spectrum $g_\alpha(Y)$ and the corresponding unwarped spectrum X is

minimized. Optimization is achieved by a full search in the interval of warping factors ranging from 0.8 to 1.2, where 1 corresponds to no warping,

$$\hat{\alpha} = argmin_{\alpha} \frac{1}{N} \sum_{i=1}^{N} \left(X_i - g_{\alpha}(Y_i) \right)^2 \tag{4.1}$$

- A linear interpolation method is introduced in order to obtain the new values of the unwarped reference spectrum to the new corresponding warped frequency values.
- The spectrum is warped according to the optimal warping factor $\hat{\alpha}$.

  Based on the evaluated optimal warping factors, we compare warping factors and spectral distances before and after frequency warping for different phonemic groups.



(a)

Figure 2. Optimal Warping Factors for the phonemes for the reference (only the male) and mapped (only the female).

In Figure 2, the optimal factors averaged for the reference (only the male one) relative to mapped (only the female) speakers are shown for phonemes. The warping factors were computed as described in the Section 4.1 and averaged over all appeared segments. From this Figure, we can conclude that there is a dependence between phonemes and warping factors.

4.2 MONOPARAMETRIC WARPING.

Frequency Warping is implemented by re-sampling the spectral envelope at linearly and nonlinearly frequency indices, i.e.

1. Linear

$$g_\alpha : \omega \rightarrow \tilde{\omega} = \alpha \cdot \omega \tag{4.2}$$

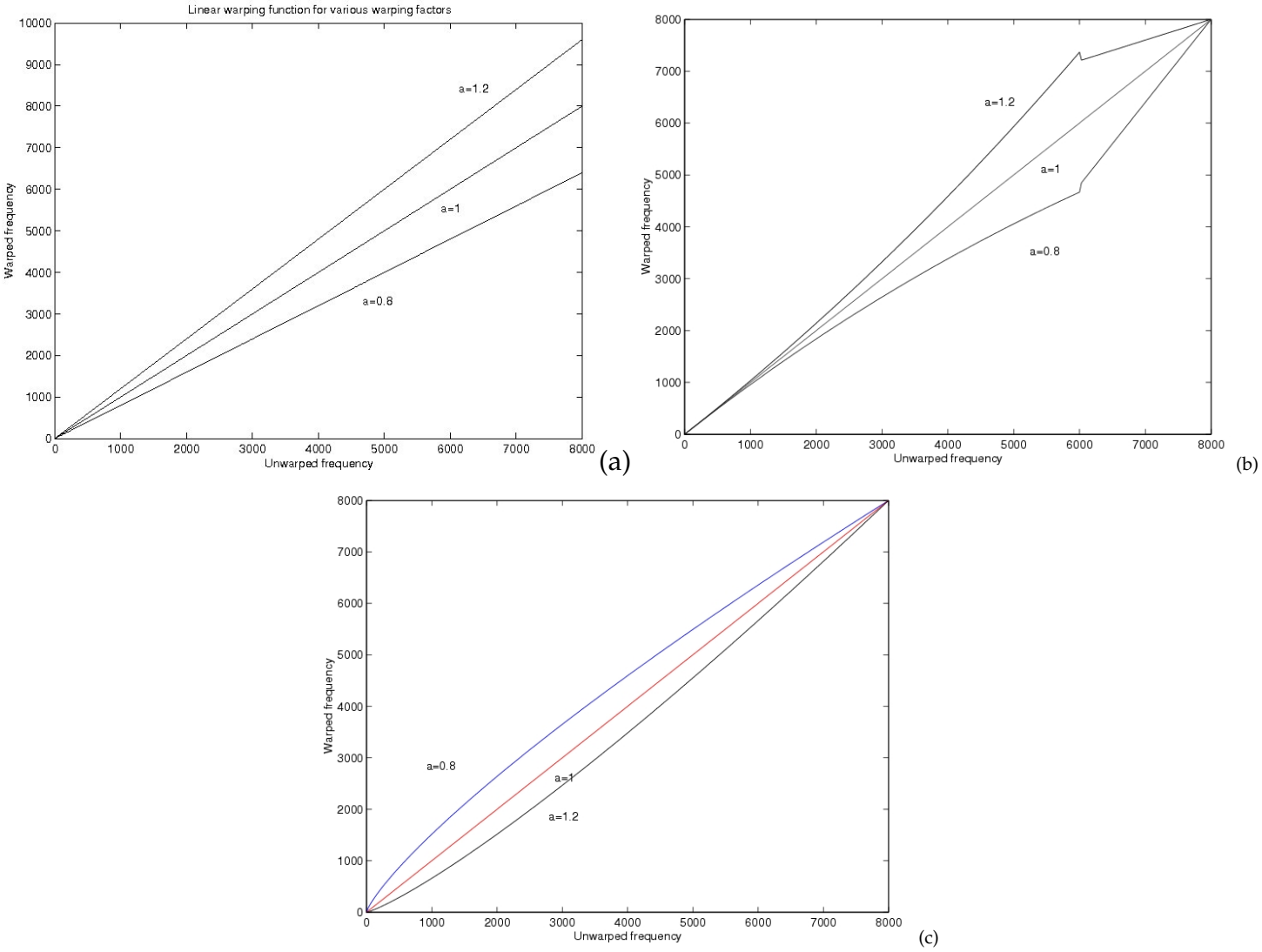


Figure 3. Monoparametric Warping functions (a) Linear, (b) PieceWise NonLinear and (c) Power for α=0.8, α=1.0 and α=1.2.

2. Piecewise Nonlinear[23]

$$g_\alpha : \omega \rightarrow \tilde{\omega} = \begin{cases} \alpha^{\frac{3\cdot\omega}{2\cdot\omega_N}} & \omega < \omega_1 \\ \beta \cdot \omega + \gamma & \omega > \omega_1 \end{cases} \tag{4.3}$$

where ω_N is equal to Nyquist Frequency, α is the warping factor and ω_1 is equal to 3 kHz. The coefficients β and γ are set to compensate for the bandwidth mismatch after warp according to the following two equations,

$$\alpha^{\frac{3\cdot\omega_1}{2\cdot\omega_N}} = \beta\cdot\omega_1 + \gamma$$

$$\omega_N = \beta\cdot\omega_N + \gamma$$

3. Power [14]

$$g_\alpha : \omega \rightarrow \tilde{\omega} = \omega_N \cdot \left(\frac{\omega}{\omega_N}\right)^\alpha \tag{4.4}$$

where ω_N is equal to Nyquist frequency.

At Figures 3a, 3b and 3c we can see the Linear, PieceWise-NonLinear and Power warping functions correspondingly for the lower ($\alpha = 0.8$), upper ($\alpha = 1.2$) limits of the values that the warping factor can take during the grid search and for no warping ($\alpha = 1$).

4.3 MULTI-PARAMETRIC FREQUENCY WARPING

Next, we try to improve VTLN performance by exploring alternative piecewise linear frequency warping strategies related with the simple linear frequency warping. For this purpose, bi-parametric and four-parametric frequency warping algorithms are proposed. For the case of bi-parametric warping algorithms, we evaluate different warping factors (α_L and α_H) for the low ($f < 3$ kHz) and high ($f \geq 3$ kHz) frequencies, correspondingly. More in detail, after the computation of the optimal warping factor $\hat{\alpha}$ of 4.1 under the constraints $|\alpha_L - \hat{\alpha}| \leq 0.04$ and $|\alpha_H - \hat{\alpha}| \leq 0.04$, a full search mode over the candidate piecewise linear warping functions will provide the optimal pair (α_L,α_H) of frequency warping factors. On this full search, the step will be equal to 0.02 and, as a result of this, the candidate functions are 25 (5^2).

An example of this procedure is presented in Figure 4 a. At this example, the optimal factor is equal to 1.06 (solid line). After that, the optimal factor for the lower than 3 kHz frequencies ($f < 3$ kHz) is equal to 1.04 and the optimal factor for the greater than 3 kHz and lower than 8 kHz ($3 < f < 8$ kHz) frequencies is equal to 1.08 (dotted line).

The procedure for the case of four-parametric is identical, only with the difference that the optimal warping factors are four corresponding to the frequency ranges, 0-1500 Hz, 1500-3000 Hz, 3000-4500 Hz and 4500-8000 Hz. The full search for the evaluation of the four optimal factors is over the 625 (5^4) different warping functions.

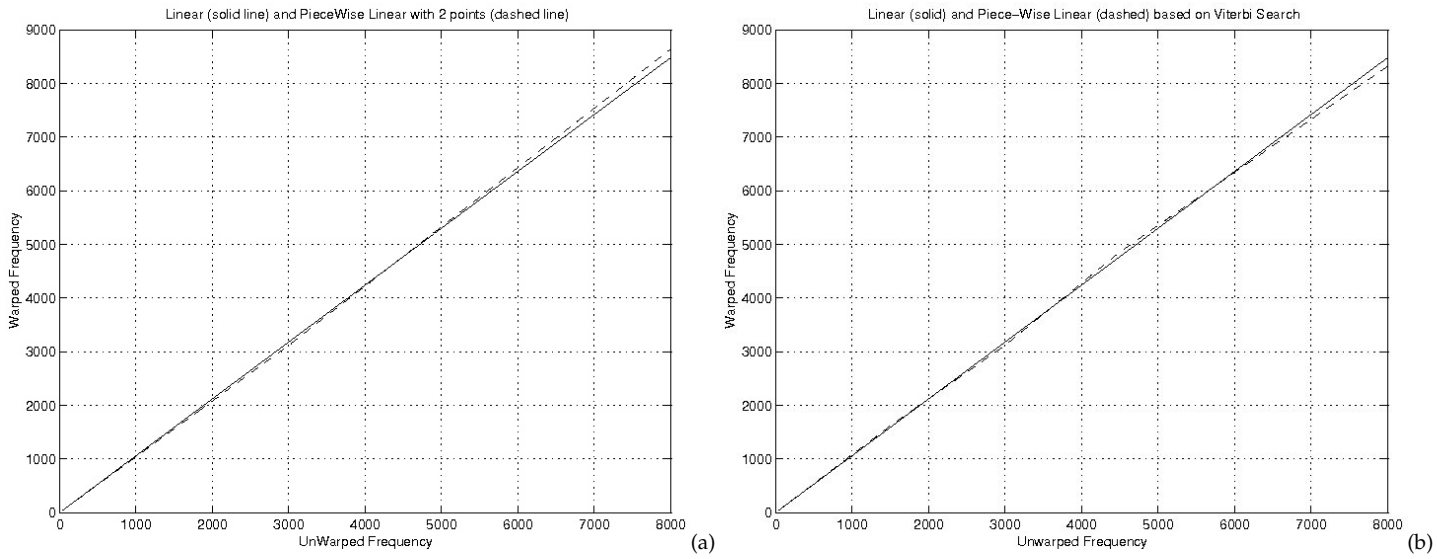


Figure 4. MultiParametric Warping Functions with: (a) two parameters and (b) four parameters.

An example of this procedure is presented in Figure 4b. The optimal, running all over the frequency range, factor is equal to 1.06 (solid line). Around this optimal value and under the same, as in two-parameters case, constraints, the optimal warping factor for the lower than 1500 Hz frequencies ($f < 1500$ Hz) is equal to 1.08, the optimal factor for the frequencies between 1500 and 3000 Hz ($1500 < f < 3000$ Hz) is equal to 1.04, the optimal factor for the frequencies between 3000 and 4500 Hz ($3000 < f < 4500$ Hz) is equal to 1.08 and the optimal factor for the frequencies between 4500 and 8000 Hz ($4500 < f < 8000$ Hz) is equal to 1.04 (dotted line).

4.4 ML ESTIMATION OF SPECTRAL BIAS.

To further minimize the spectral distance between the reference and mapped spectrums a spectral bias is computed. In general, one could hypothesize different formulations for the bias. However, the spectral bias **b** is chosen to be linear,i.e.

$$b = c \cdot f + e \tag{4.5}$$

The bias estimation, i.e the parameters c and e, is performed by Maximum Likelihood, as follows:

- For each phoneme and speaker and for the middle frame of the utterance, the average spectral envelope is computed,

- Using the ML algorithm, we obtain the two parameters (**c** and **e**),

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - Y_i - c\dot{f}_i - e\right)^2$$

Using the ML algorithm, we obtain the two parameters (**c** and **e**),

$$\frac{dMSE}{dc} = 0 \Rightarrow \frac{2}{N}\sum_{i=1}^{N}\Big((X_i - Y_i - cf - e)(-f)\Big) = 0$$

$$\sum_{i=1}^{N} f\Big((X_i - Y_i - cf - e)\Big) = 0 \Rightarrow$$

$$\sum_{i=1}^{N} f\Big((X_i - Y_i - e)\Big) = \sum_{i=1}^{N}\Big(cf^2\Big) \tag{4.6}$$

Keeping, for the moment the equation 4.6, we will obtain, through ML, the coefficient e of the linear bias:

$$\frac{dMSE}{de} = 0 \Rightarrow \frac{2}{N}\sum_{i=1}^{N}\Big((X_i - Y_i - cf - e)(-1)\Big) = 0 \Rightarrow$$

$$\sum_{i=1}^{N}\Big((X_i - Y_i - cf - e)\Big) = 0 \Rightarrow$$

$$\sum_{i=1}^{N}(X_i - Y_i) + \sum_{i=1}^{N}(-cf) + \sum_{i=1}^{N}(-e) = 0 \Rightarrow \tag{4.7}$$

Through the evaluation of the spectral envelope, we removed the DC value from the two spectrums. This is the reason why we didn't choose a constant bias to add to the unwarped mapped spectrum. That fact modifies the equation 4.7,

$$\sum_{i=1}^{N}(e) = 0 + \sum_{i=1}^{N}(-cf) \Rightarrow$$

$$e = \frac{1}{N}\sum_{i=1}^{N}(-cf) \Rightarrow$$

$$e = -c \cdot \frac{1}{N}\sum_{i=1}^{N} f \tag{4.8}$$

Replacing in equation 4.6 the equation 4.8,

$$(4.6) \Rightarrow \sum_{i=1}^{N} f\Big((X_i - Y_i - (-c \cdot \frac{1}{N}\sum_{i=1}^{N} f))\Big) = \sum_{i=1}^{N}\Big(cf^2\Big)$$

$$\sum_{i=1}^{N} f(X_i - Y_i) + \sum_{i=1}^{N} f(c \cdot \frac{1}{N}\sum_{i=1}^{N} f) = c \cdot \sum_{i=1}^{N} f^2$$

$$\sum_{i=1}^{N} f(X_i - Y_i) = c\Big(\sum_{i=1}^{N} f^2 - \sum_{i=1}^{N} f(\frac{1}{N}\sum_{i=1}^{N} f)\Big)$$

Finally, the two coefficients of the linear bias are equal to:

$$c = \frac{\sum_{i=1}^{N} f(X_i - Y_i)}{\Big(\sum_{i=1}^{N} f^2 - \sum_{i=1}^{N} f(\frac{1}{N}\sum_{i=1}^{N} f)\Big)} \tag{4.9}$$

and

$$e = -c \cdot \frac{1}{N}\sum_{i=1}^{N} f \tag{4.10}$$

- The extracted linear bias is added to the unwarped mapped spectrum.
- The frequency warping method, taking account of the modified mapped spectrum, continues as it is described at Section 4.1.

A typical example of the above procedure is presented in the Figures 5a, 5b and 5c. The reference speaker is the *mhito* (male) and the mapping is the *mdhlo* (male). The selected phoneme from these two speakers is the "eh" phoneme. In the Figures 5a and 5b are presented the reference's and mapping's unwarped spectrums and in the Figure 5c is presented the extracted linear bias for this case. Finally, in Figure 5d is the new, after the addition of the bias, "mapping" spectrum, i.e. the unwarped "mapping" spectrum added with the extracted linear bias.

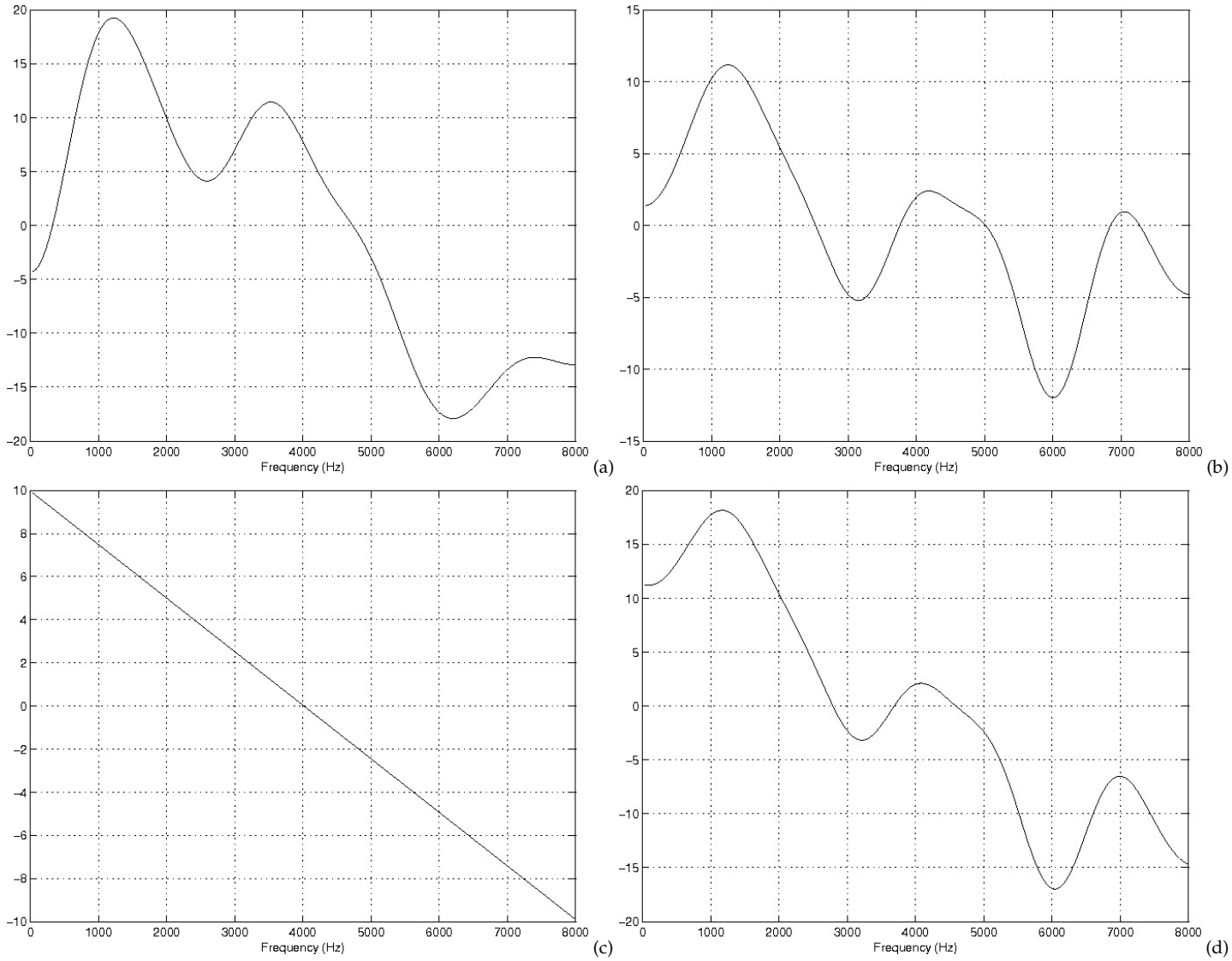


Figure 5. Spectrums: (a) Reference speaker's spectrum (b) Mapped speaker's spectrum (c) Maximum Likelihood Estimated Linear Bias which is added to the mapping spectrum before the optimal warping factor estimation process. (d) After the addition mapping spectrum.

4.5 RESULTS

In Figure 6, the average (over all speakers) MSE between the "reference" and "mapped" speakers before and after the frequency warping is computed for various phonemes. Additionally to that, the intra-speaker variability for each phoneme is shown. At the same Figure, the upper point at lines corresponds to the average (over all speakers) MSE between the "reference" and "mapped" speakers

before the frequency warping and the lower one corresponds to the average (over all speakers) MSE between the "reference" and "mapped" speakers after the frequency warping. More specifically, at Figure 6a, the dot line corresponds to the effect at MSE that the linear warping provides. The middle line corresponds to the Piecewise-Nonlinear and the solid one to the Power function. At Figure 6b, the dot line corresponds to the effect at MSE that the mono-parametric warping provides, the middle line corresponds to the bi-parametric warping and the solid one to the four-parametric function. At Figure 6c, we present the influence of the mono, bi and four parametric warping combined with the added linear bias.

Examining the Figure 6a, we may conclude that the simple linear frequency warping is shown to be efficient in reducing acoustic mismatch for most phonemic classes. Also, the piecewise nonlinear warping function is also comparatively efficient with the linear case, especially for the vowels. Note the relatively large distance reduction for vowels, glides and the small reduction for fricatives /f/ and /z/. Finally, for the /ao/,/eh/,/ih/,/aw/ and /r/ phonemes, the intra speaker variability is, at the worst case, achieved.

Based on the Figure 6b, we may see clearly the further minimization to the spectral distance reduction provided by the two and four-pieces warping functions. However, we don't achieve the averaged spectral distance be equal to or less than the intra speaker variability for more phonemes, than those we had on the simple linear case.

In Figure 6c, we present the results of the bias evaluation and insertion before the frequency warping process. Examining this figure, we conclude that the extra parameter (linear bias) plays an important role on the further reduction of the spectral distance. The intra-speaker variability is achieved for the majority of the phonemes.

4.6 SPEAKER DEPENDENT VARIABILITY

Beyond the MSE averaged over all speakers, we will examine the speaker variability averaged over all phonemes. As it is already been said, the reference speakers are 8, 4 males and 4 females. It is important to check the effects of the frequency warping for the reference speakers and all the phonemes that they have pronounced.

Looking the Figures 7a and 7b, we notice that after multi parametric normalization, the intra-speaker variability levels are achieved for most speakers.

4.7 PERCENT DISTANCE REDUCTION.

In the current section, percent distance reduction, when scaling factors are computed on an per phoneme basis, is investigated. In Figures 8a, 8b and 8c, the mean and standard deviation of the percent reduction in spectral distance between phonetic utterances of reference and mapping speakers due to warping is shown

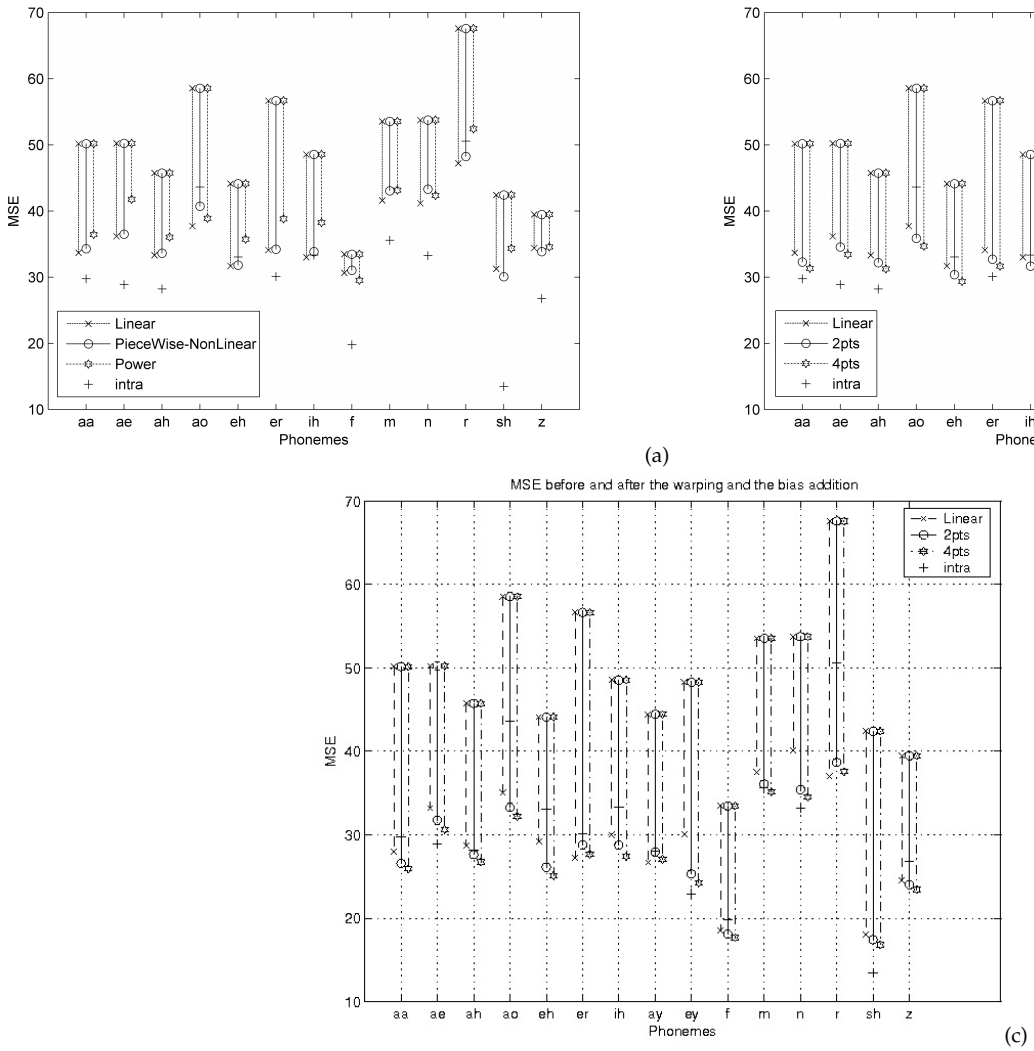


Figure 6. Intra-speaker variability (+) and averaged MSE between reference and mapped speakers (male and female) before and after warping: (a) linear, piecewise-nonlinear and power warping functions (b) bi-parametric (2pts) and four-parametric (4pts) warping (c) bi-parametric (2pts) and four-parametric (4pts) warping and the linear bias addition.

for monoparametic and multiparametric with two and four parameters warping cases.

In these plots the scaling factor and distance reduction are computed for each phonemic instance. Note that mono- and multi- parametric frequency warping are effective normalization procedures.

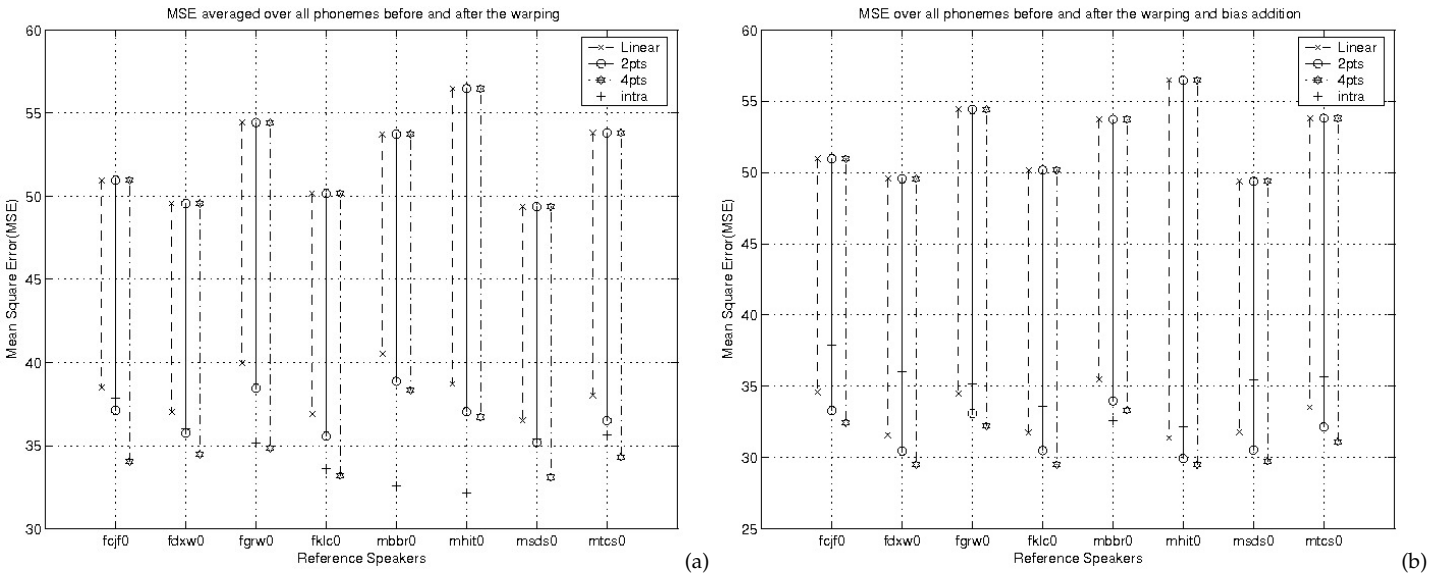


Figure 7. (a) Averaged over all phonemes Mean Square Error before and after the linear, bi-parametric (2pts) and four-parametric (4pts) warping for all reference speakers and intra-speaker variability (+). (b) Averaged over all phonemes MSE before and after the linear, bi-parametric (2pts) and four-parametric (4pts) warping and bias addition for all reference speakers and intra-speaker variability (+).

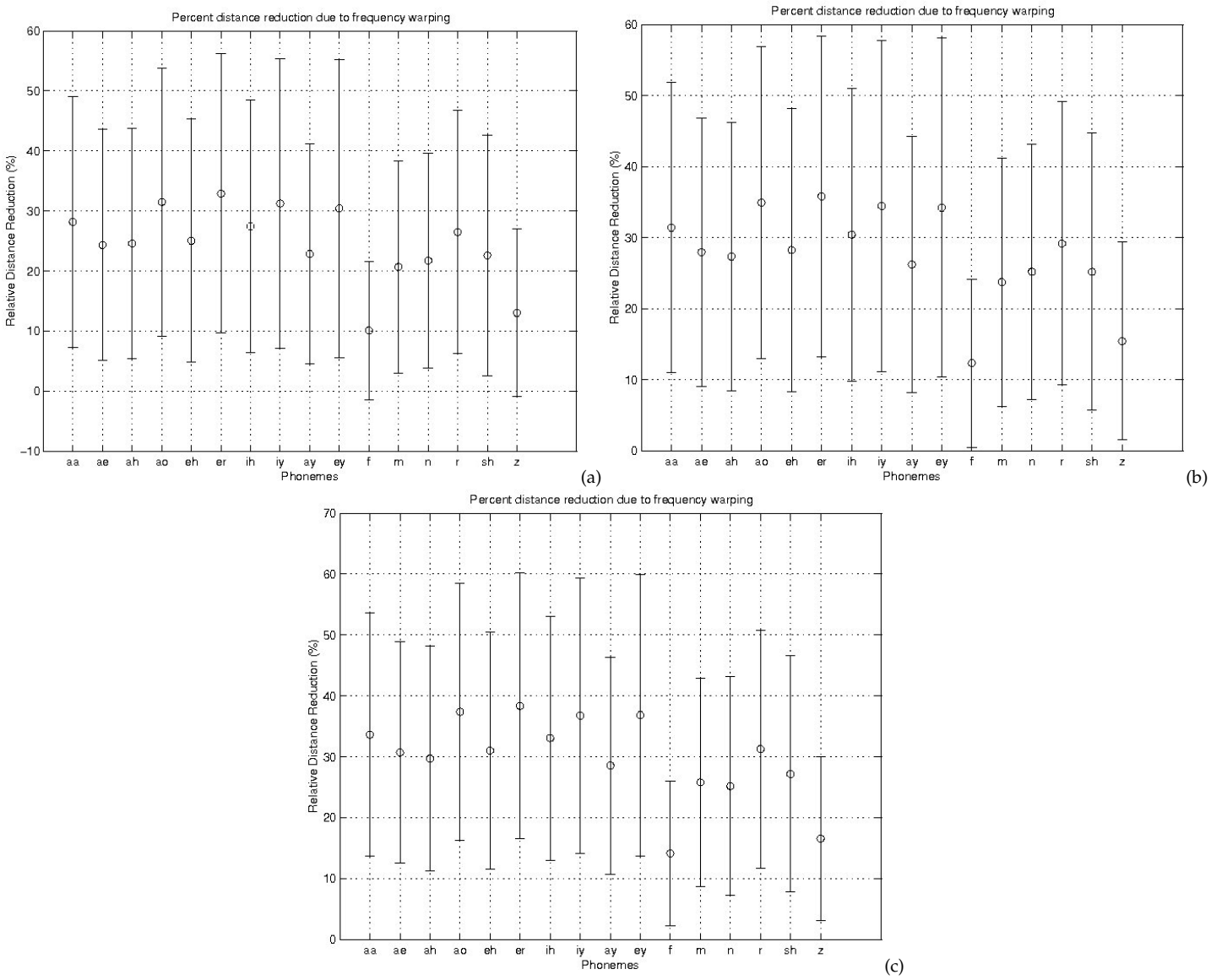


Figure 8. Percent distance reduction due to frequency warping when scaling factors and distance reduction are computed on an per utterance basis. Mean and standard deviation of distance reduction (error bars) are displayed for: (a) linear, (b) bi-parametric (2pts) and (c) four-parametric (4pts) case.

5

REGION-BASED VTLN

Vocal tract length normalization is a model based normalization scheme that relies, in particular, on the size of the vocal tract. The vocal tract, i.e. the position and shape of the human organs determine the generated sound. Formant center frequencies of the speech signal are inverse proportional to the length of the vocal tract. Since the vocal tract length varies from about 13 cm for female to over 18 cm for male speakers, there are systematic inter-speaker variations of formant frequencies by up to 25% [12].

In this section, we present the Region- based VTLN method (R-VTLN). This method, firstly, categorizes the testing utterance's frames into regions and then evaluates region-specific spectral warping functions and factors using an ML criterion. The proposed method is implemented during the testing procedure.

5.1 FRAME SEGMENTATION.

The first step in R-VTLN method is the classification of the testing utterance's frames into regions. Two independent algorithms are examined. More specifically, each utterance's cepstral vectors may be classified through two unsupervised algorithms.

One of the proposed unnsupervised frame categorization algorithm is based on KMeans algorithm. KMeans is one of the simplest unsupervised learning algorithms which solve clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids shroud be placed in a good way because different location causes different result. So, the better choice is to place them as much as possible far away from each other. In the case of this thesis, for the appropriate initialization of our algorithm, we take a part of AURORA4's training dataset (examining carefully to be gender-independent) and we pass it through the KMeans algorithm in order to extract the appropriate centroids. The population of these centroids is equal to the number of the regions we choose to categorize the utterance's frames. Feature extraction, consisting of a Hamming window 25 ms and a frame update of 10 ms, results in the standard 13 dimensional cepstrum Mel Cepstrum Cepstral Coefficients (MFCC) consists of the 12 static coefficients including zero coefficient. Taking the first and second order coefficients, we result to the 39-dimensional MFCC. These 39-dimensional cepstral coefficients are the input of the KMeans algorithm. The algorithm aims at minimizing an objective function, in this case the cepstral distance. Our proposed algorithm (*KM* henceforth) is composed of

the following steps:

1. Based on a subset of training data (equal to fifty training utterances) extract P centroids (equal to the population of regions). These points represent the initial group centroids.
2. Assign each frame F to the region c with the closest centroid, based on the minimization of the cepstral distance D, i.e.

$$D = \sum_{i=1}^{N} |F_i - C_i^j|^2 \tag{5.1}$$

   where N=39 (dimension of MFCC), C_i^j is the i'th coefficient of the centroids obtained by the j'th iteration of the K-Means algorithm and F_i is the i'th each frame's cepstral coefficient.
3. When all frames are assigned, recalculate the positions of the centroids.
4. Repeat Steps 2 and 3 until no significant change to the centroids.

5.1.1 *Unsupervised Phonetic-Class Assignment.*

An alternative way to classificate the frames into regions is based on the conclusions of the dependence between the phonemes-warping which is already defined at Section 4.3 and on the phonemic-label recognized labels from a first recognition pass. More specifically, a first recognition pass provides us the correspondence between frames and phonemes for the recognized labels. At this point, we return to the conclusions extracted at Section 4.3 and, based on them, we supervisedly categorize the phonemes of the monophone lexicon into regions. At Chapter 6, where results will be presented, we will also present at Tables the exact phonemic categorization for the various population of regions.

5.1.2 *Constraints*

It is important to note that frames's population included in each region should be capable in order to determine the optimal factors and functions. That means that the population of the frames should be greater than a, determined by us, threshold. In both of the categorization methods, (*KM* and *PhCat*), we define this threshold be at least one hundred (100) frames per region. For the achievement of this constaint, the KM algorithm is well initialized (which means that the initial centres are well estimated) and at *PhCat* case, the phonemes are divided by us in such way that we include a capable number of frames in each region.

Both of the two proposed categorization algorithms's output is a mapping O between the L frames and their region index sequence R, $O : l \rightarrow r$. Following the frame classification algorithm, a median filtering is applied on the sequence R. Median filtering applies one-dimensional median filter to vector sequence R. At a first step, the values in the window are sorted. After that, for each frame k in the window, what it is selected as its region index $R'(k)$, when n is odd, is the median of $L(k-(n-1)/2 : k+(n-1)/2)$. The filtered sequence R' is the same length as sequence R. Median filtering is used to smooth these inherently noisy frame assignments, based on the continuity criterion. At Figures 9a and 9b, we can see the before and after smoothing region index sequence for the two regions case.

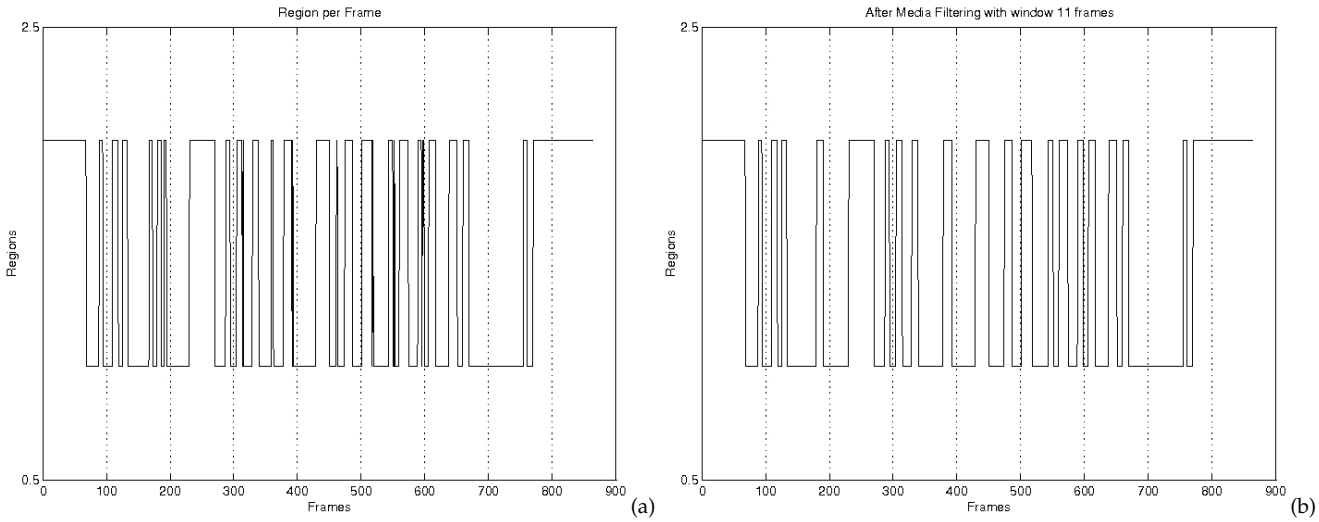


Figure 9. Region Index Sequence of 440c0204.wv1 utterance (a) Before the smoothing and (b) After the smoothing.

At Figure 9, we may see the region index sequence before and after the smoothing for the utterance "440c0204.wv1" spoken by the testing speaker 440 and sampled at 8 kHz when the number of regions are two (2). At axis Y we have the regions while at axis X we have the sequence of frames. After the smoothing, we may see that the noisy passing from the one region to another is smoothed.

5.2 WARPING PROCEDURE

After the frame categorization, the spectral coefficients corresponding to each region are warped according to one of the M factors and one of the N functions g. This results to a multi-dimensional warping process, which obtains the P optimal factors and functions for each region by maximizing the likelihood of the warped vectors with respect to the transcriptions from the first pass W and the

unnormalized HMM λ,

$$\hat{\vec{\alpha}}, \hat{\vec{g}} = argmax_{\vec{\alpha},\vec{g}} P(X^{\vec{\alpha},\vec{g}}|\lambda, W) \tag{5.2}$$

where

$$\vec{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ ... \\ \alpha_P \end{bmatrix}, \vec{g} = \begin{bmatrix} g_1 \\ g_2 \\ ... \\ g_P \end{bmatrix} \tag{5.3}$$

The optimal parameters can be determined by

- an exhaustive search over factors and functions for all the regions simultaneously (referred as *Sim* henceforth). That means a full optimization of the warping factor, given the word sequence from a first recognition pass. During this optimization, the search over the ensembles of M factors α and N functions is taking place simultaneously for all the regions. Through a grid search over the various factors and functions, the final optimal for each region warping factor and function is estimated.
- To avoid the full optimization, the warping factor α may be determined by evaluating the optimal factors and functions independently for each region (referred as *Sep* henceforth). More in detail, an iterative procedure is taking place. The total iterations are equal to the region population. At each one of these iterations, through a search over the factors and functions separately for each region, we evaluate the optimal factor and function for this region while the other regions's frames are warped linearly by the global factor α_{glb},

$$\alpha_k = \begin{cases} \alpha_m & \text{if } k = r, \\ \hat{\alpha}_{glb} & \text{if } k \neq r \end{cases} \tag{5.4}$$

$$g_k = \begin{cases} g_n & \text{if } k = r, \\ Linear & \text{if } k \neq r \end{cases} \tag{5.5}$$

The complexity of the proposed method at Section 5.2 *Sim*, is equal to $M^P \cdot N$, where M is the length of the ensemble of the factors, P is the region population and the N is the length of the candidate functions's ensemble (linear and piece-wise linear). The problem of raised complexity at *Sim* case can be solved by the *Sep* method. The proposed method *Sep* has complexity similar to that of Lee-Rose's method, equal to $M \cdot P \cdot N$.

5.3 REGION-BASED VTLN IN RECOGNITION.

During recognition the following two pass strategy is followed:

- through a first recognition pass, a transcription W is obtained using the unwarped sequence of cepstral vectors X and the unnormalized model λ. Beyond the transcription which is required for the alignment and, of course, is not given at testing procedure, we extract the phoneme-level labels which will be used at *PhCat* method.
- Through the *KM* or the *PhCat* method, the utterance's frames are categorized into P regions.
- Perform a forced alignment procedure for each considered warping factor and function. We choose as optimal this set of factor and function that maximizes the conditional probability given the preliminary transcription and the unnormalized acoustic model,

$$\hat{\vec{\alpha}}, \hat{\vec{g}} = argmax_{\vec{\alpha},\vec{g}} P(X^{\vec{\alpha},\vec{g}}|\lambda, W) \tag{5.6}$$

- The warped with $\hat{\vec{\alpha}}$ and $\hat{\vec{g}}$ sequence $X^{\hat{\vec{\alpha}},\hat{\vec{g}}}$ is decoded in order to obtain the final recognition result.

Recognition test results on AURORA4 and CHIMP corpus for two-pass VTLN and for R-VTLN method are summarized in Chapter 6.

6 EXPERIMENTAL RESULTS

6.1 EVALUATION SETUP

For the evaluation of the normalization method R-VTLN, speech recognition tests were performed. The speech recognition experiments involved recognizing the test material as if it was unknown speech, and then comparing the recognition result to the known transcription. For the evaluation, the parameter measured in the experiments is the recognition word accuracy.

6.2 EXPERIMENTAL PREPARATION

Feature extraction consisted of a Hamming window 25 ms and a frame update of 10 ms, resulting in the standard 39 dimensional cepstrum coefficients consisted of the zero coefficient, the static and their delta and the delta-delta coefficients. The KMeans was used also to compute ceptral distances during the frame classification procedure (KM). For the smoothing constraint, the length of the window in frames is equal to nine (9). For all experiments the optimum, for each region, warping factor is obtained by searching between $0.88 \leq \alpha_m \leq 1.12$ with step 0.02, which leads to an ensemble of 13 factors. The candidate warping functions are the Linear and the Piecewise-Linear, which are shown in Fig. 10. For the Linear case, the lower limit f_L is equal to 125 Hz (f_L = 125 Hz) for both cases of AURORA4 - 8 kHz and 16 kHz sampling frequency, while the upper limit f_U is equal to 3980 Hz (f_U = 3980 Hz) for the AURORA4 - 16 kHz and f_U is equal to 7960 Hz (f_U = 7960 Hz) for the AURORA4 - 16 kHz sampling frequency. .

All the techniques were evaluated with the speech materials, which are already described at Chapter 9. More specifically, for the case of AURORA4, the clean training set of AURORA4 (7138 utterances, 128.294 words) is used to train the unnormalized acoustic models. Two clean testing subsets of 330 utterances from 8 speakers (total 5353 words) sampled at 8 and 16 kHz are used as evaluation sets. More details about the training and testing databases are provided at the Appendix 2. Speaker independent models are trained. A continuous speech recognition task requires a language model, so, for the case of AURORA4, a bigram language model is used.

The tests were evaluated using monophone HMM models. Five-state per phoneme models with one Gaussian per state were trained including silence models with three states per phone. HMM silent-pause model consist of a single state and single Gaussian per state is trained.

At first evaluation, the regions in which each utterance's frames are categorized

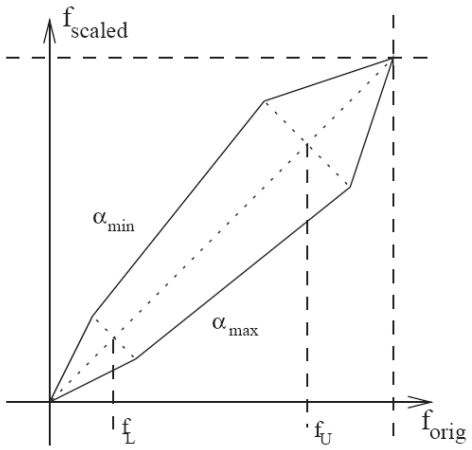


Figure 10. Linear and Piecewise Linear warping functions.

are two. At two regions case and for the *PhCat* method, the phonemes of the vocabulary are divided according to the following Table 1.

Phonemes per Region	
Region-1	Region-2
aa,ae	g,b,jh, ch
ah,ao	l, d,m, dh
eh,er	n, f,ng, k
ey,iy	r, p,hh, s
ih,ow	v, sh,w, t
oy, y	z, th
uh,uw	zh, sil,sp

Table 1. Phonemes Per Region for the Two Regions Case.

Examining carefully the Table 1, we may see that the vowels are assigned to one region and all the others monophones, including silence and silent-pause are assigned to the other region. This categorization is based to the extracted from the study at Chapter 4 conclusions. At that study, we may see that the vowels display similar behaviour in the minimization of the MSE. Beyond that, previous studies have presented the great importance that the vowels play during the extraction

of the warping factor [20]. This is the reason why we discriminate one region only for the vowels and, as a first step, a second one for all the other phonemes including silence.

Extending the two regions case, the regions that each utterance's frames are categorized are increasing to three. At Table 2 for the *PhCat* method, we may see the categorization of the phonemes at three regions. The silence is excluded from the second class and a third class is including it. Usually, the silence is a turbulence factor, this is a reason why we exclude it from the second region creating a third region with it.

Phonemes per Region (three Regions Case)		
Region-1	Region-2	Region-3
aa,ae,eh,er,ih,ow,uh	g,b,jh, ch,n, f,ng, k,v, sh,w, t	sil
ah,ao,ey,iy,oy, y,uw	l, d,m, dh,r, p,hh, s,z, th,zh	sp

Table 2. Phonemes Per Region for the Three Regions Case.

Extending the three regions case to five regions, we divide the monophones according to the Table 3.

Phonemes per Region (Five Regions)	
Region 1	ey,ay,aa,ae,iy,ih, ah
Region 2	uh,uw, aw,ao,ow, oy,er, eh
Region 3	jh, ch,dh, sh,th, zh,m,n, hh, f
Region 4	g, k, w, d, b, p, t
Region 5	l,z, j, s, sh,r, sil, sp

Table 3. Phonemes Per Region for the Five Regions Case.

6.3 EXPERIMENTAL RESULTS: AURORA4 - 8 KHZ.

At Table 4, the experimental results of the proposed R-VTLN for the two, three and five regions are presented.

The VTLN two-pass method proposed by [12] improves significantly the baseline performance (4.9%). Our proposed methods (both of *R-VTLN KM-Sim* and *R-VTLN PhCat-Sim*) come to improve the accuracy further by 3% (161 words) and 2.9% the two-pass method. For the case of *R-VTLN KM-Sim* we have the greatest

Regions	2	3	5
Baseline	48.3		
VTLN (Lee-Rose)	53.2		
R-VTLN KM-Sim	56.2		
R-VTLN PhCat-Sim	55.6		
R-VTLN KM-Sep	56.1	55.7	55.4
R-VTLN PhCat-Sep	55.8	55.6	55.4

Table 4. Word accuracy results (%) evaluated on clean test set of AURORA4. The sampling frequency is equal to 8 kHz.

improvement in performance always compared with the [12] proposed method (3.4%, 169 words). Given the small difference in performance between the *Sim* and *Sep* algorithms, only results for the *Sep* algorithm are shown for three and five regions.

Results degrade somewhat when using five regions. This could be due to the lack of adequate data to estimate multiple parameters (a single utterance is used here for warping factors and functions estimation) and decreasing returns from using multiple regions.

The distribution of the optimal factors and functions is another issue of this thesis. At the following Figures, we will present selected factors's and functions's distributions for the methods described in this thesis.

More in detail, in Figure 11 the distributions of the difference between the region-based warping factors and the global warping factor $\hat{\alpha}_{glb}$ are shown for the *KM-Sim* method. As expected the region factors lie around the global optimal factor and take lower values for the first region and somewhat higher for the second region. It has to be mentioned that the frames corresponding to "start" and "end" of the utterance are categorized to the first region. Probably, the turbulence appeared at first region is a result of this fact. The distribution of the optimal factors and functions when the method is *PhCat* (Figure 21) at Apeendix 1 improves the turbulence of the distribution of region 3 which includes the "sil" and "sp" frames.

Despite of that, the fact that the optimal factor for the second region lies very close to the global factor is something that we expected because our method comes to extend the unique global factor provided by Lee-Rose's method[12]. Most probably, that region includes the frames which corresponds to vowels. The initial centers of the KMeans algorithm are constant during the entire method's progress and provided by an initial KMeans. This KMeans categorizes a subset of training frames chosen from several training speakers and provides the initial

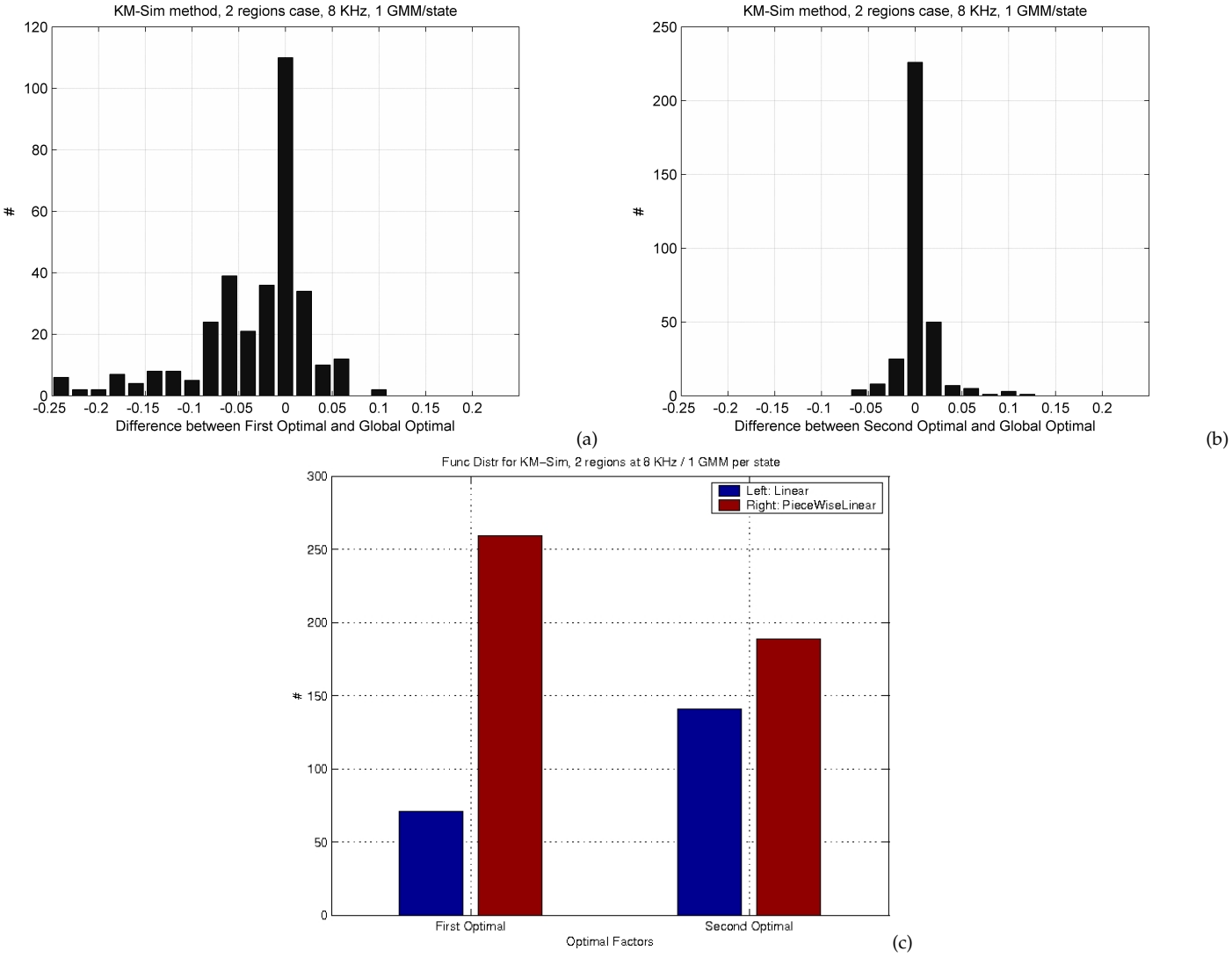


Figure 11. Distribution of the Difference Between the First Optimal Factor (a), the Second (b) with the Global Factor α_{glb} For the Two Regions Case and the *KM-Sim* method. Also, the distribution of the chosen as optimal warping function for the two regions (c).

centers which after it is used at our method.

As far as it concerns the optimal per region functions, we can see from the Figure 11c that in both regions, the piecewise linear dominates to the linear. However, for the second region the contribution of the two functions is about the same. At Appendix 1, we can see further the distributions of *KM-Sep* and *PhCat* methods when the regions are two.

Beyond the two regions case, we study the distributions of the differences between the optimal per region factors and the global factor for the three regions case. In Figure 12 and for the *KM-Sep* method, we can see these differences concerning the three optimal factors and the corresponding optimal functions. From the Figure 12a, we conclude that the first factor takes values mostly equal with the global factor. At the same time, the same turbulence as in the case of the two regions is presented. We have to mention that the "start" and the "end" of the utterance is categorized again to the first region. The other two factors lie around the global one, taking values higher than the global one (the second optimal) and lower than the global one (the third optimal) Figure 12b and 12c correspondingly).

Looking carefully the Figure 12d which corresponds to the functions, we conclude that for the first region, the piecewise linear dominates to the linear while to the rest regions the contribution of these two regions is almost shared.

In Figure 13, we can study the distributions of the differences between the optimal factors and the global one when the regions are five. The first factor is corresponded to the "start" and "end" of the utterance and lies around the global optimal. The other optimal lie around the global taking values lower (second optimal and fifth) and higher (third and fourth optimal).

Beyond the contribution of the factors, we will examine the contribution that the functions provides to our method. At Table 5, we can see the results from the two-pass method (which uses the linear warping function) compared with the R-VTLN when only linear function is used and when the method chooses from the ensemble of the functions (linear and piecewise linear). The regions are Two and the Gaussian Mixture Per State is equal to one.

Examining carefully the results at Table 5, we may see that we have a great improvement from baseline to Lee-Rose's method. This improvement is extending when our method is inserted using only one (the same with Lee-Rose's method) warping function, the linear one. When the ensemble of functions encloses both the linear and the piecewise linear functions, the improvement becomes further and promising for further improvement if the ensemble is further enhanced.

Next we investigate if these improvements hold for HMM models of increasing complexity. Results are presented in terms of word accuracy for the two regions case and the KM-Sep method for the AURORA 4 task at Table 6, when the sampling frequency is 8 kHz.

When the Gaussians per state are increased, as expected, baseline performance also increases significantly. At the same time the relative improvement of VTLN

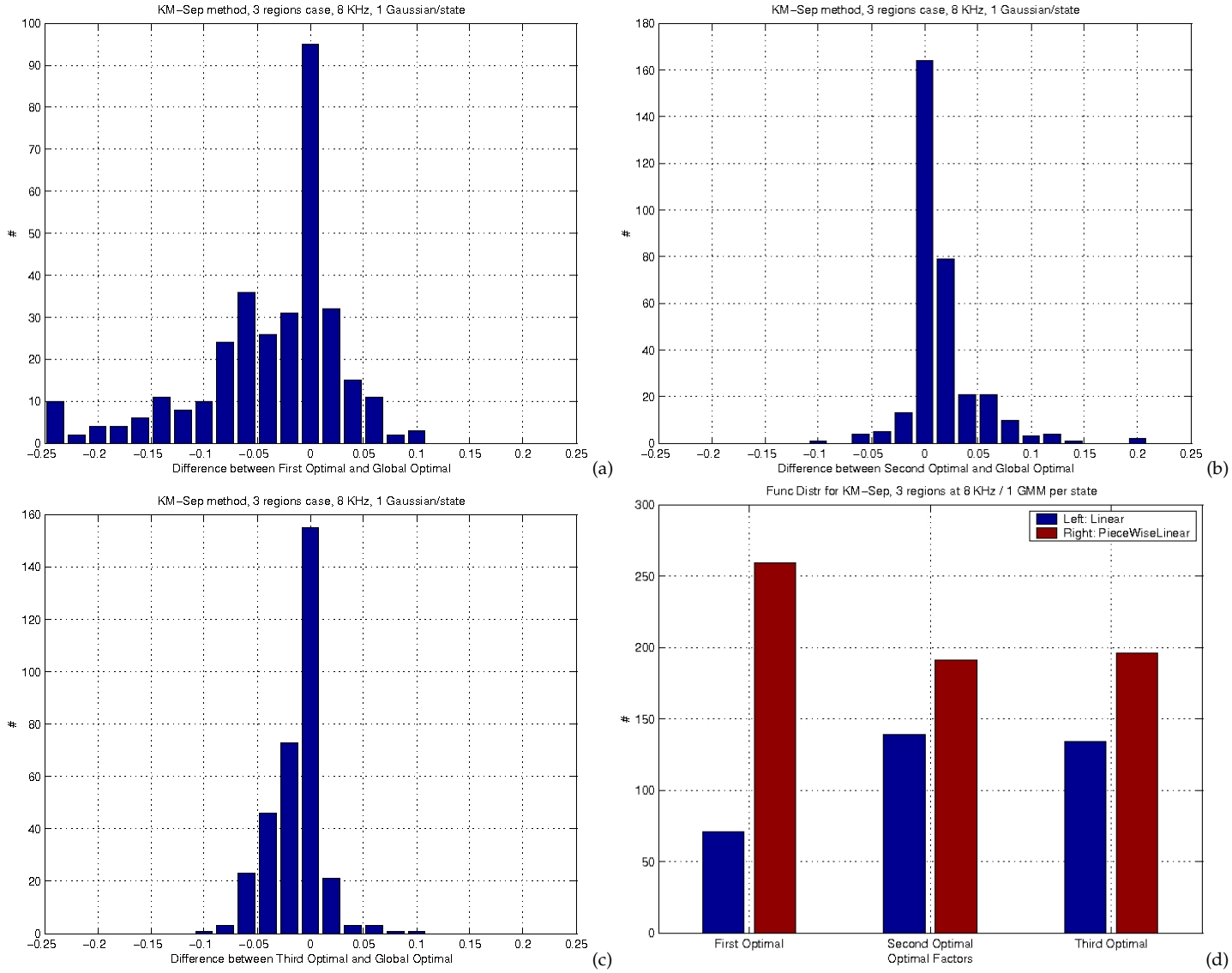


Figure 12. Optimal Factors Distribution For the Three Regions Case ((a), (b) and (c)) and the *KM-Sep* method. Also, the distribution of the optimal warping functions for each region (d).

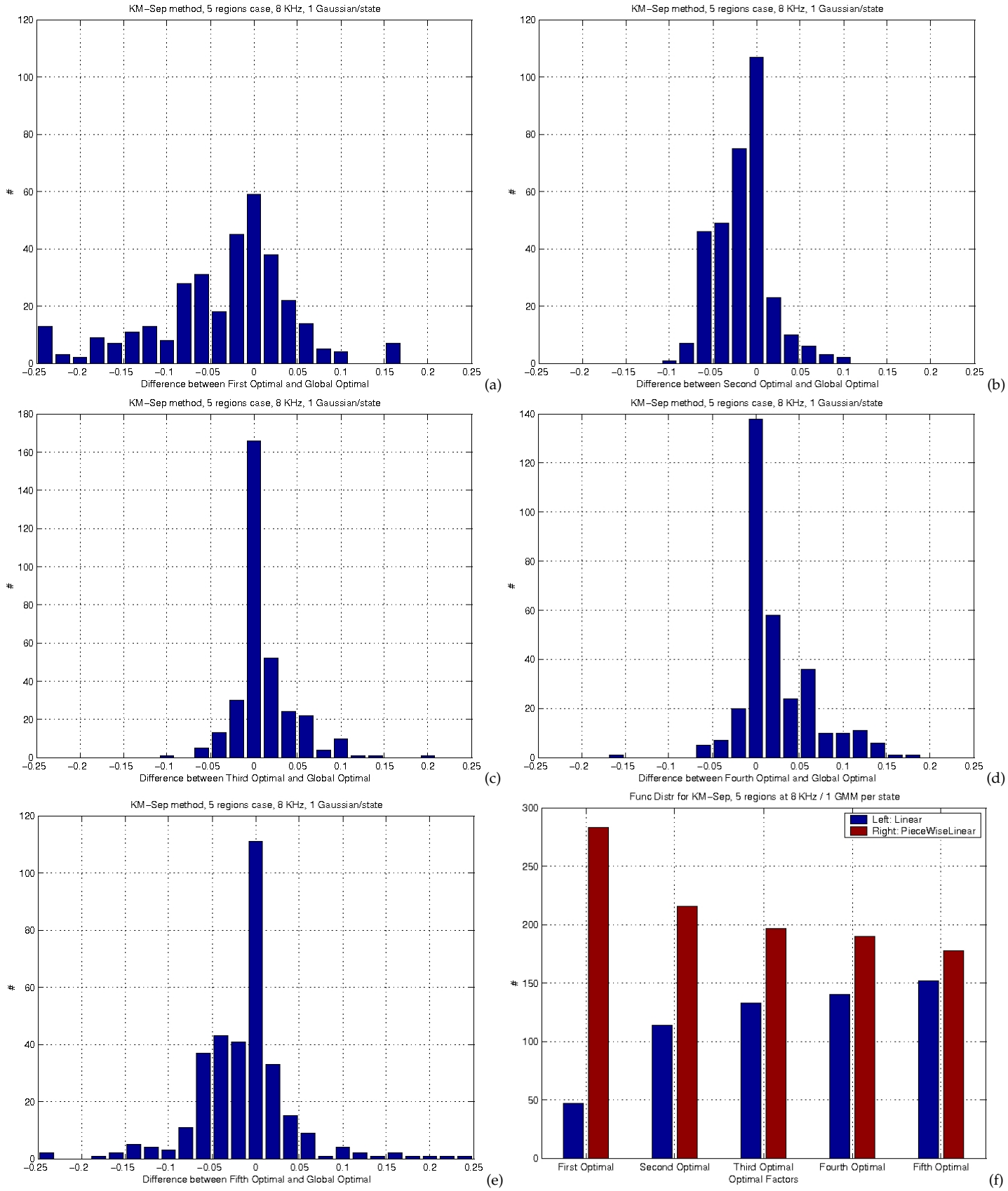


Figure 13. Distribution of the Difference Between the First Optimal Factor (a), the Second (b), the Third (c), the Fourth (d) and the Fifth (e) with the Global Factor α_{glb} For the Five Regions Case and the *KM-Sep* method. Also, the distribution of the chosen as optimal warping function For the Five Regions Case (f).

Two Regions Case - 1 Gaussian HMM	
Baseline	48.3
VTLN (Lee-Rose)	53.1
R-VTLN KM-Sim (Only Linear Fun)	55.7
R-VTLN KM-Sim (Linear or Piecewise Linear)	56.2

Table 5. Word accuracy results (%) when the ensemble of functions at R-VTLN encloses only the Linear Warping Function and when R-VTLN chooses the optimal warping function from a function ensemble which encloses the Linear and Piecewise Linear warping Functions.

Two Regions Case			
GMM per State	*1*	*3*	*8*
Baseline	48.3	55.4	56.4
VTLN (Lee-Rose)	53.1	60.1	60.7
R-VTLN KM-Sim	56.2	63.1	63.5
R-VTLN KM-Sep	56.1	63.5	63.9
R-VTLN PhCat-Sep	55.8	63.6	64.2

Table 6. Word accuracy results (%) versus the number of Gaussian Mixtures per State on monophones HMM.The sampling frequency is equal to 8 kHz.

over baseline decreases. However, the improvement of R-VTLN over VTLN remains consistently the same. For eight Gaussians per state the improvement of VTLN over baseline is comparable to the improvement of R-VTLN over VTLN. The distributions of the optimal factors and functions (Figures 23 and Figure 24 at Appendix 1) remains the same as in 1 GMM/state.

At Figure 14, we may see the averaged over the sentences WACC versus the length of the sentences for the case of *KM-Sep* method evaluated at AURORA4 - 8 kHz. Examining carefully this Figure, we may see the positive contribution of the method *KM-Sep*. However, as long as the number of regions arises, which means that the parameters needed to be evaluated arises, the WACC decreases. The danger of overfitting is becoming more obvious while we increase the number of regions.

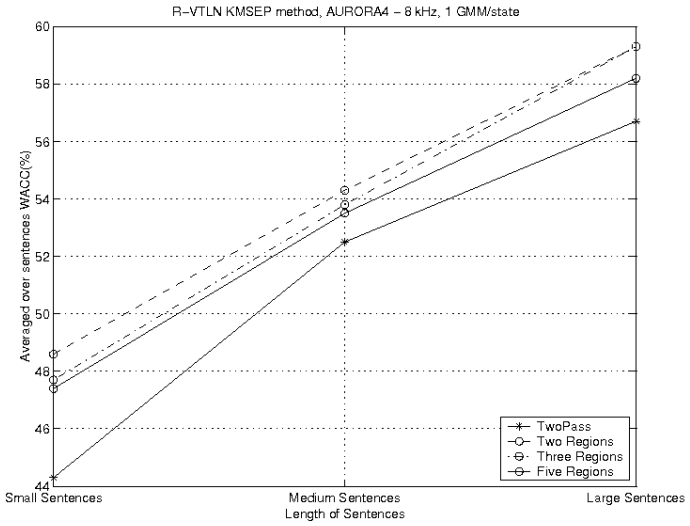


Figure 14. Distribution of the averaged over sentences WACC versus the length of the testing sentences. The results are based on the KM-Sep method evaluated on AURORA4 - 8 kHz with 1 GMM/state.

6.4 AURORA4 - 16 KHZ

Beyond the 8 kHz sampling frequency, we evaluate the R-VTLN method to training and testing AURORA4's testing subset sampled at 16 kHz. At Table 7, the experimental results of the proposed R-VTLN for the two regions case are presented.

Regions	2	3	5
Baseline	55.2		
VTLN (Lee-Rose)	57.4		
R-VTLN KM-Sim	60.9		
R-VTLN PhCat-Sim	60.6		
R-VTLN KM-Sep	60.8	60.8	60.5
R-VTLN PhCat-Sep	60.7	60.7	60.8

Table 7. Word accuracy results (%) evaluated on clean test set of AURORA4.

Examining carefully the Table 7, we may see that the baseline performance increases significantly. At the same time the relative improvement of VTLN over baseline is about 2.2%. However, the improvement of R-VTLN over VTLN remains

as in the 8 kHz case (above 3%).

Next we investigate if these improvements hold for HMM models of increasing complexity. Results are presented in terms of word accuracy for the two regions case and the KM-Sep method for the AURORA4 task at Table 8 for one, three and eight Gaussians per state.

Two Regions Case			
GMM per State	*1*	*3*	*8*
Baseline	55.2	61.9	63.1
VTLN (Lee-Rose)	57.4	65.1	66.3
R-VTLN KM-Sep	60.8	67.4	68.9
R-VTLN PhCat-Sep	60.7	67.4	68.7

Table 8. Word accuracy results (%) versus the number of Gaussian Mixtures per State on monophones HMM.

When the Gaussians per state are increased, as expected, baseline performance also increases significantly. At the same time the relative improvement of VTLN over baseline increases. However, the improvement of R-VTLN over VTLN decreases. For eight Gaussians per state the improvement of VTLN over baseline is comparable to the improvement of R-VTLN over VTLN. The distributions of the optimal factors and functions (Figures 30 and Figure 31 at Appendix 1) remains the same as in 1 GMM/state.

We will examine the distribution of the optimal factors and functions. At Figure 15, we may see the two optimal factors and functions for the *KM-Sim* method and the two regions case. Both of the optimal take values around the global factor. At 16 kHz case, we see that the first region's factor takes lower values, while the second region's factor takes values clearly above the global factor. For the function case, we can see that the piecewise linear function dominates to the linear in both regions.

Extending to three regions case and for the *KM-Sep* method, at Figure 16, we may see the distribution of the optimal factors and functions. The conclusions remains the same as in two regions case.

For the regions case, we choose to present the distribution of the *PhCat-Sep* based on the supervised categorization of Table 3. All the factors lie around the global taking lower or higher values. The domination of the piecewise linear function versus the linear remains.

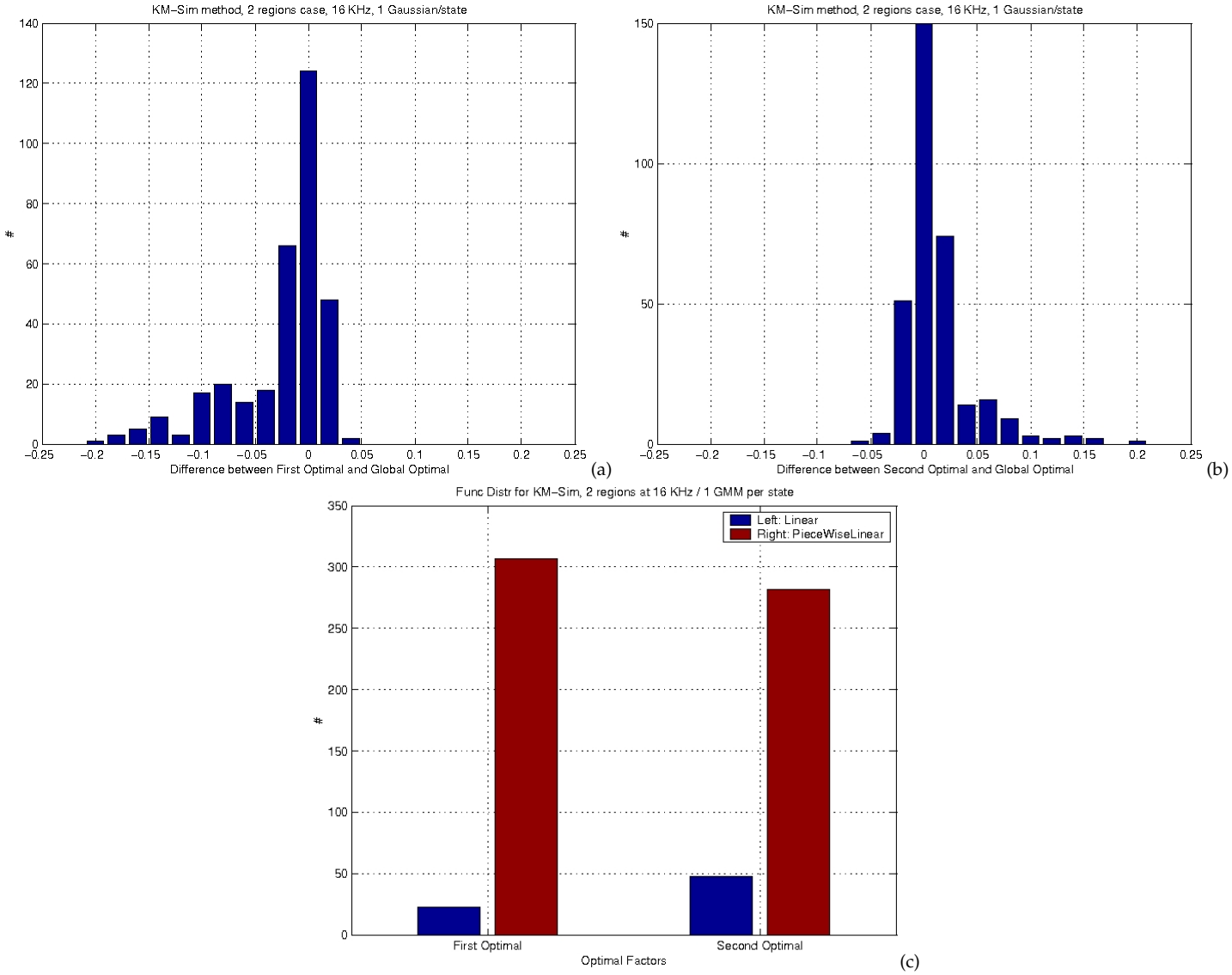


Figure 15. Distribution of the Difference Between the First Optimal Factor (a), the Second (b) with the Global Factor α_{glb} For the Two Regions Case and the *KM-Sim* method. Also, the distribution of the chosen as optimal warping function for the two regions (c).

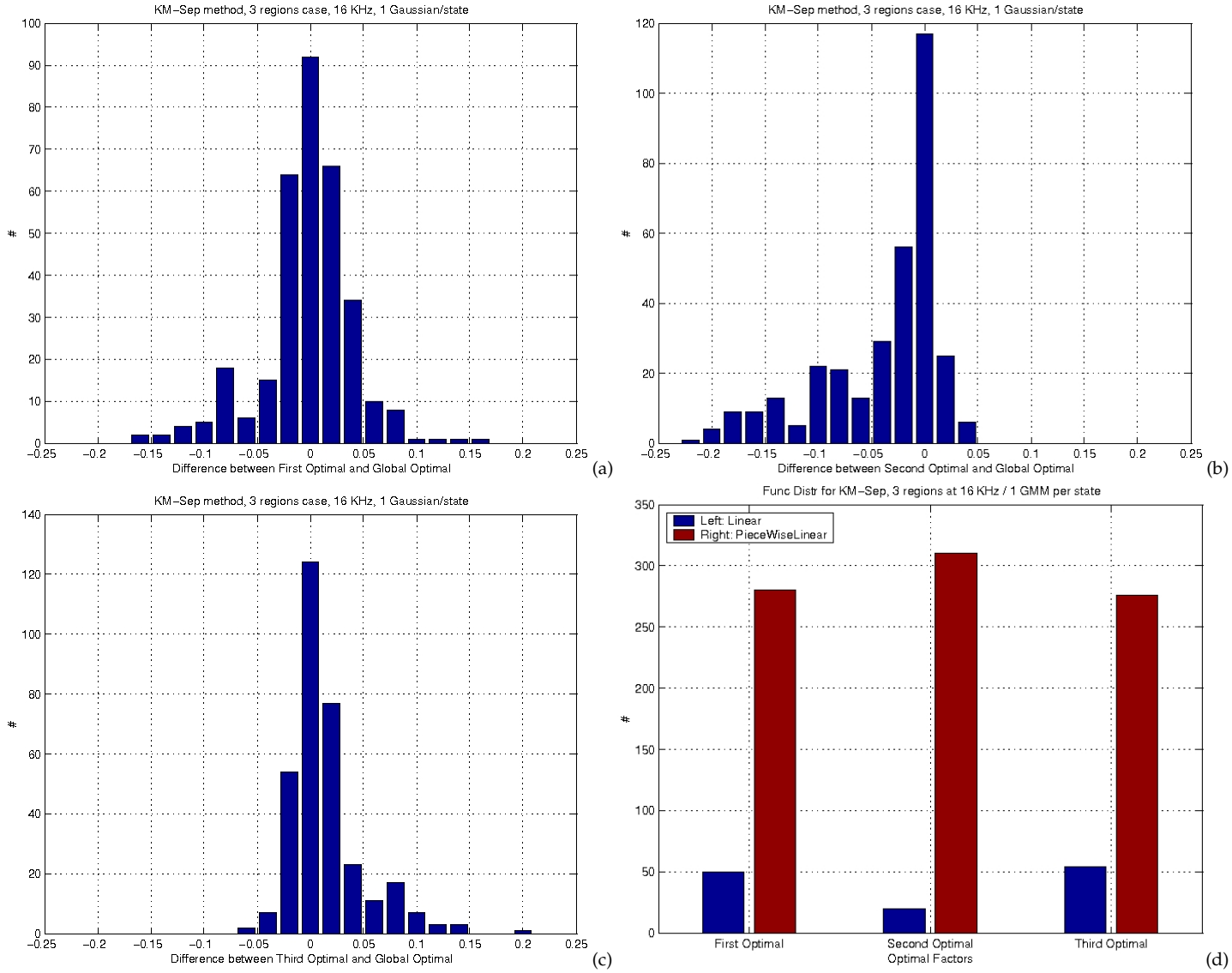


Figure 16. Optimal Factors Distribution For the Three Regions Case ((a), (b) and (c)) and the *KM-Sep* method. Also, the distribution of the optimal warping functions for each region (d).

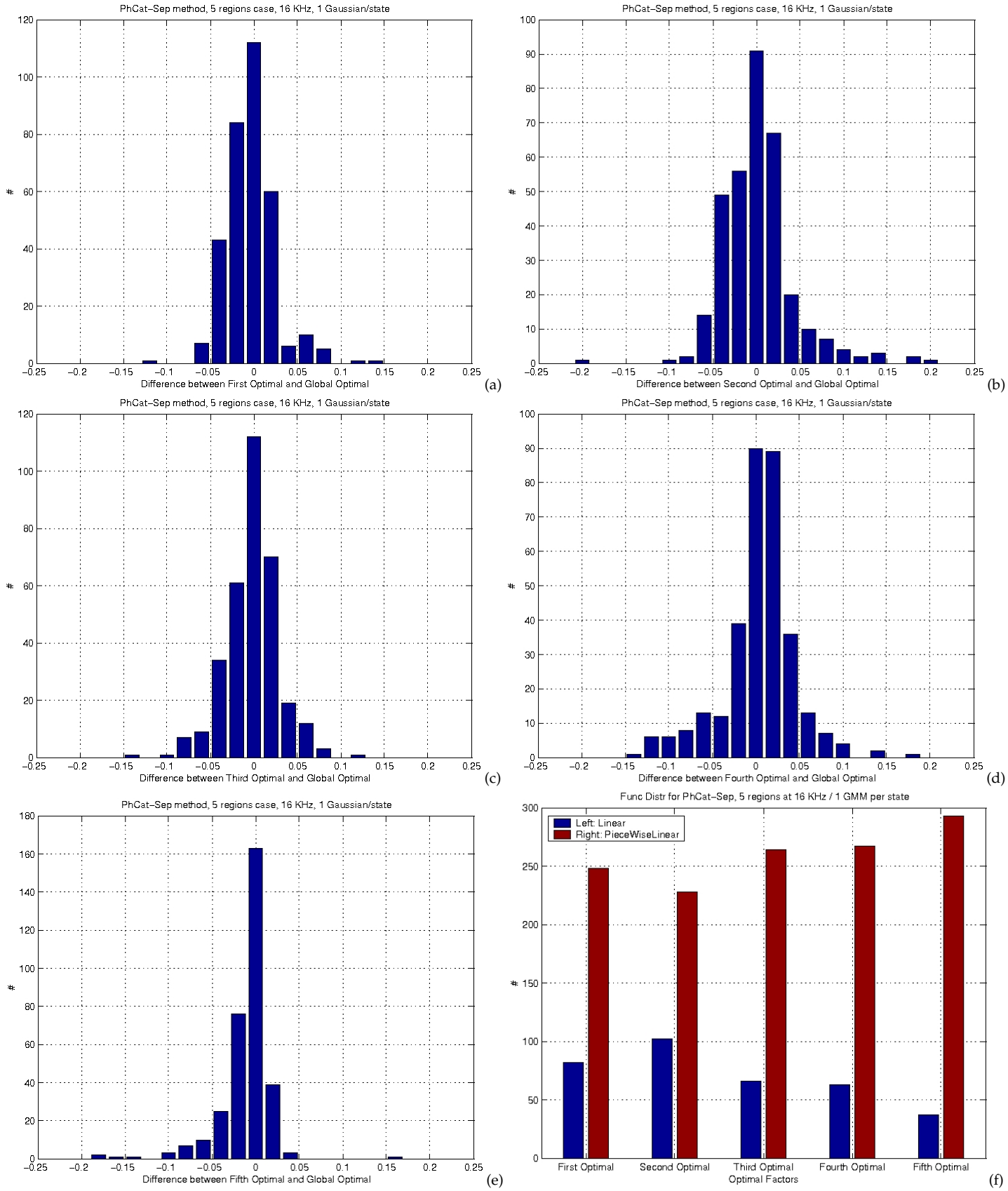


Figure 17. Optimal Factors Distribution For the Five Regions Case and the *PhCat-Sep* method.

7

CONCLUSIONS-FUTURE WORK

7.1 CONCLUSIONS.

In this thesis, a Region Based Vocal Tract Length Normalization as a means to improve recognition results was studied. Today, most normalization algorithms operate only with one parameter (the warping factor α). At frequency domain, the warping is applied on the entire utterance. These approaches does not take account of the phonemes's behavior to the warping process.

First of all, the between phonemes and warping process dependence is examined. We have shown quantitatively the dependence between frequency warping functions and phones. The behavior of the phones during the warping process is different for the various phones. This is the main reason why in this thesis, we made a step forward and we investigated the idea of warping an entire uterance with locally constrained warping factors and functions. We proposed a region-based VTLN algorithm. In this algorithm, at a first step utterance's frames are classified in regions through unsupervised methods. After that, a region-dependent warping factor and function are extracted based on Maximum Likelihood criterion. This method extends the unique-level method of [12] and it is implemented during recognition procedure. Because of that, this new system has been compared to the already proposed methods by Lee-Rose [12] regarding the word accuracy of their output.

R-VTLN was evaluated on AURORA4 and it was shown that significant gains over utterance-based VTLN can be achieved with a small increase in computational complexity. Among the R-VTLN variants the algorithm using k-means for frame classification and region-independent warping factor computation was shown to be competitive in both performance and computational complexity.

In order to be certain that our results are valid and that the test set that have been used is large enough to provide confident results, we performed significance tests over our results. Using the proposed method at [29], we examined the confidence that our experimental results are statistically significant and that the new proposed system has clearly improved on the baseline and already described two-pass method.

The system is implemented using information from a first recognition pass at both points. During the frame categorization and more specifically at *PhCat* method based on the phoneme-level labels which enclosed the correspondence between frames and phonemes. The same labels are used during the force segmentation at recognition procedure based on the fact that on testing procedure, the transcriptions of the testing utterances are unknown.

As the field of ASR may be considered, it is evident that locally contrained (frame level) methods will soon be a matter of the future. Combining normalization and adaptation methods shall definitely improve their performance.

7.2 FUTURE WORK.

In the future we will investigate better criteria for regions's selection. The domain will be cepstral or spectral. One proposed method may be the training of a codebook in which, at recognition procedure, the frames's categorization will be based.

Based on a training part, a codebook will be created. Consider one of these observation sequences $Y = y_1, y_2 ... y_S$, length S frames. A frame Y_s is categorized to the region $\hat{C}$ if and only if,

$$\hat{C} = argmin_{C_i} d(y_s, C_i)$$

where d is the squared Euclidean distance between the two vectors. At the same time, through a first pass from the HMM models, we evaluate each frame's likelihood $P(Y_s/\lambda)$ where λ is the Hidden Markov Model.

After the completion of the testing frames's categorization and the evaluation of their corresponding likelihood, we evaluate an averaged likelihood for each region. This averaged likelihood is equal to the mean value of the likelihoods from the frames categorized in each region. If N of S frames have been categorized to region i, the averaged likelihood is equal to :

$$L_i = \frac{1}{N} \sum_{j=1}^{N} P(Y_j/\lambda)$$

The codebook is ready to use it during the recognition procedure.

During the recognition procedure, an optimal warping factor $\hat{\alpha}$ is computed frame by frame, so that the algebraic distance between the region's averaged likelihood L_i that the warped Y_s^{α} is already categorized and each warping factor's obtained likelihood, $P(Y_s^{\alpha}/\lambda)$ is minimized. Optimization is achieved by a full search in the interval of warping factors ranging from 0.8 to 1.2, where 1 corresponds to no warping,

$$\hat{\alpha} = argmin_{\alpha} |(P(Y_s^{\alpha}/\lambda) - L_i)|$$

Based on this optimal factor, we choose the warping function with the optimal parameter(s) (one or more). One parameter means linear warping, the others means piecewise linear warping. After that, we can warp each frame with optimal factor and function. In case of multi-parametric warping function, beyond the

warping factors ((α_1,α_2,α_3,α_4) for the warping function's pieces, we can add as parameters the inflections points where the slope of the warping function changes (ω_1,ω_2, ω_3). Recognition pass with take place based on the optimally warped frames.

Also, the ensemble of the warping functions can be enriched with alternative warping functions (Piecewise Linear with More Parameters, Piecewise Nonlinear, Power) in order to model better the behavior of the phonemes at the frequency domain.

Recently, published investigations have combined normalization with adaptation methods in order to eliminate the spectral mismatch between training and testing utterances [5], .

Also, another locally constrained method can focus on state level. At this level, an optimal factor can be obtained based on the following method:

- Through a first recognition pass, we extract the optimal state level ($\hat{s}$) for each frame.
- Based on this optimal state and ML criterion, we extract an optimal factor for each frame through a search over the 13 factors (0.88 to 1.12 with step 0.02). That means that for each frame,

$$\hat{\alpha} = argmax_{\alpha} P(X^{\alpha}/\lambda, \hat{s})$$

- Second recognition pass with the warped utterances.

8

APPENDIX 1: FACTORS AND FUNCTION DISTRIBUTION.

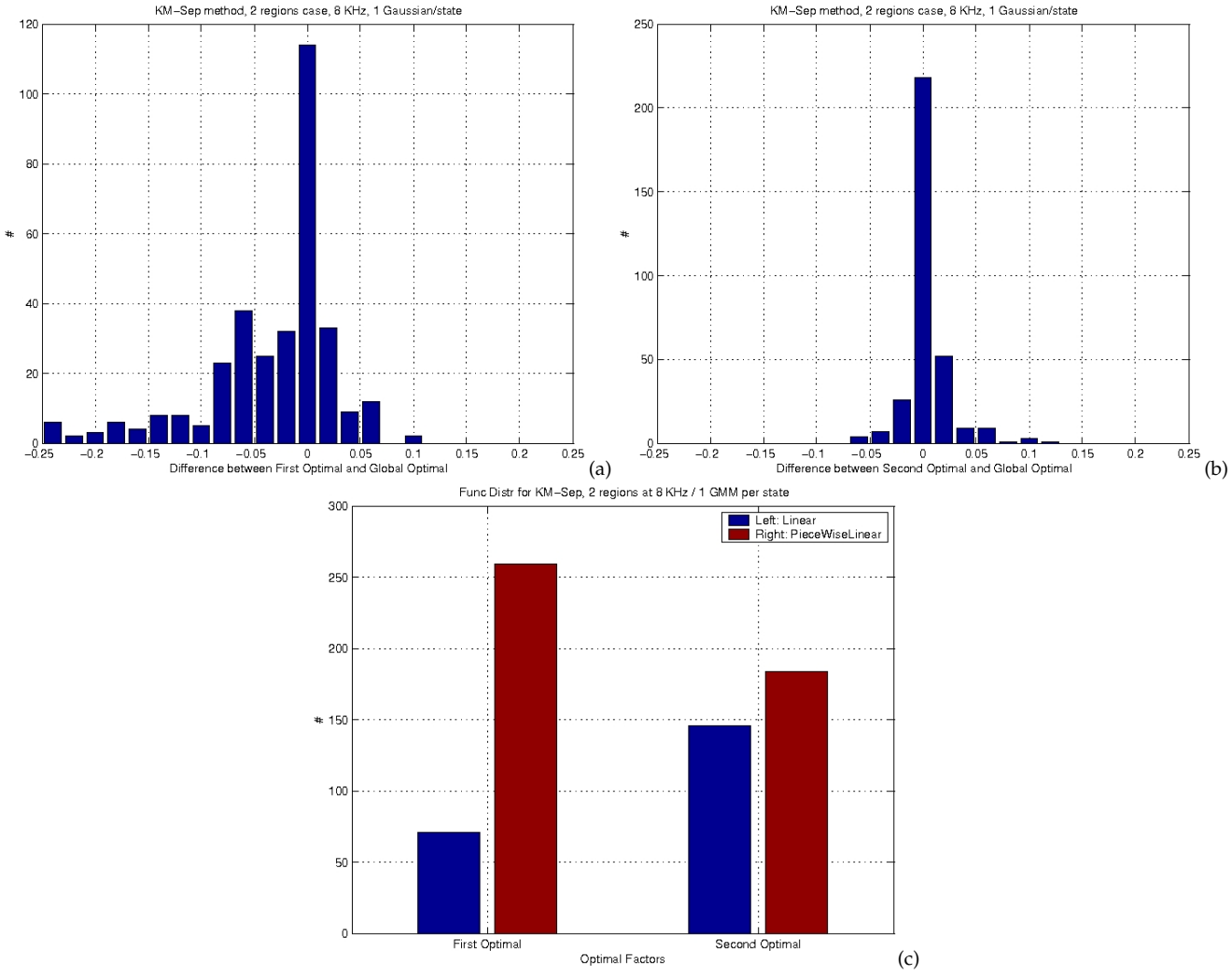


Figure 18. Distribution of the Difference Between the First Optimal Factor (a), the Second (b) with the Global Factor α_{glb} For the Two Regions Case and the *KM-Sep* method. Also, the distribution of the chosen as optimal warping function for the two regions (c).

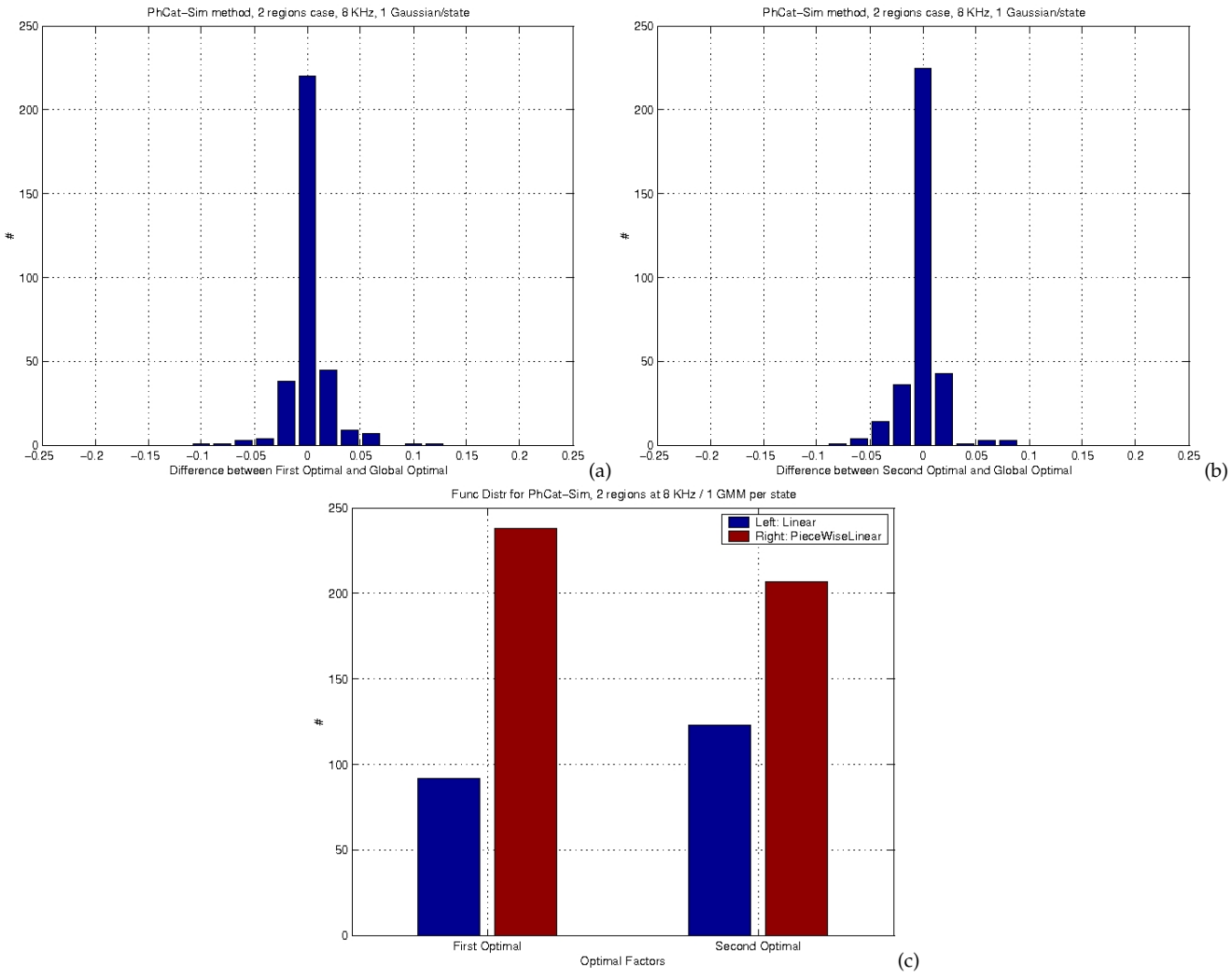


Figure 19. Distribution of the Difference Between the First Optimal Factor (a), the Second (b) with the Global Factor α_{glb} For the Two Regions Case and the *PhCat-Sim* method. Also, the distribution of the chosen as optimal warping function for the two regions (c). The sampling frequency is equal to 8 kHz.

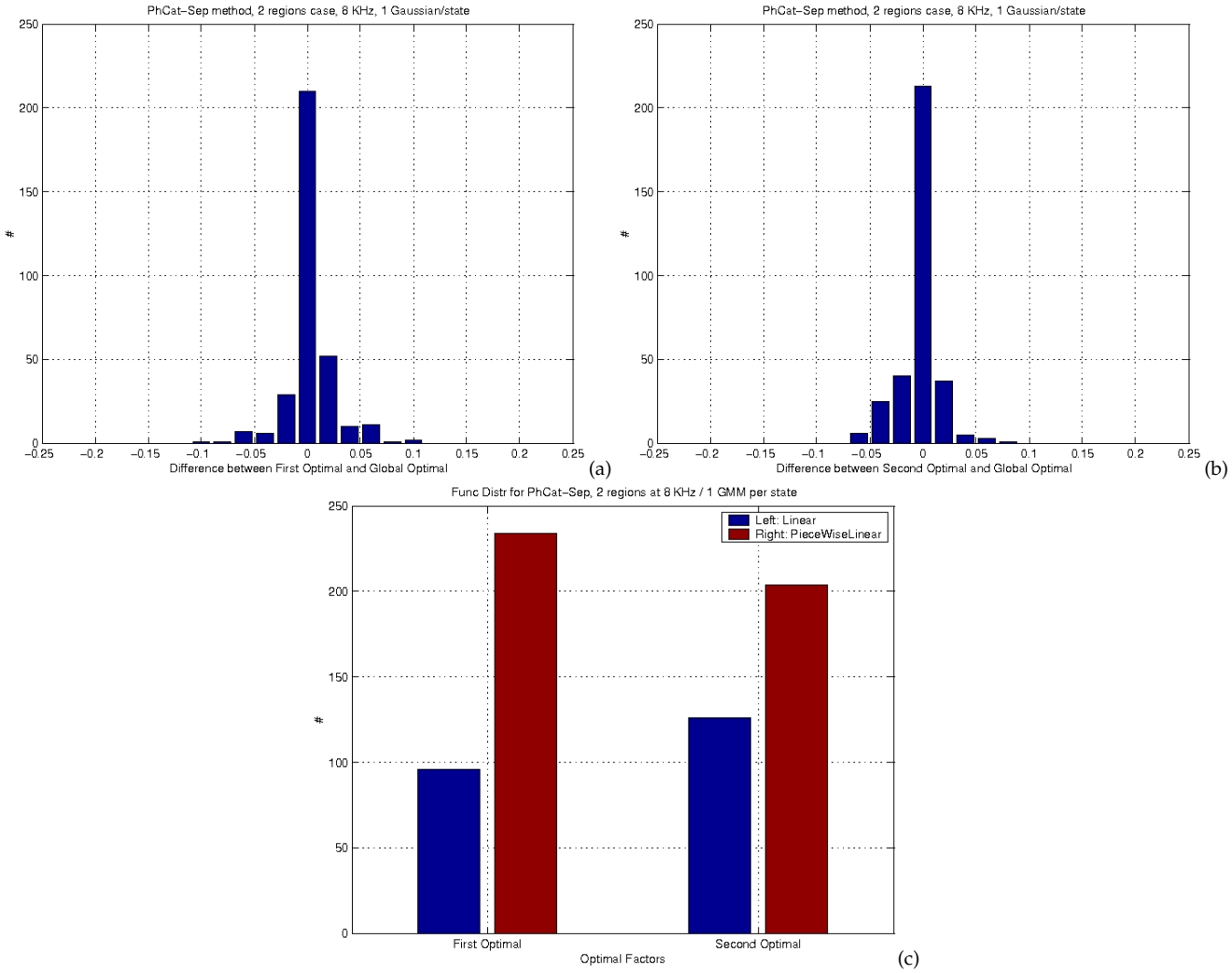


Figure 20. Distribution of the Difference Between the First Optimal Factor (a), the Second (b) with the Global Factor α_{glb} For the Two Regions Case and the *PhCat-Sep* method. Also, the distribution of the chosen as optimal warping function for the two regions (c). The sampling frequency is equal to 8 kHz.

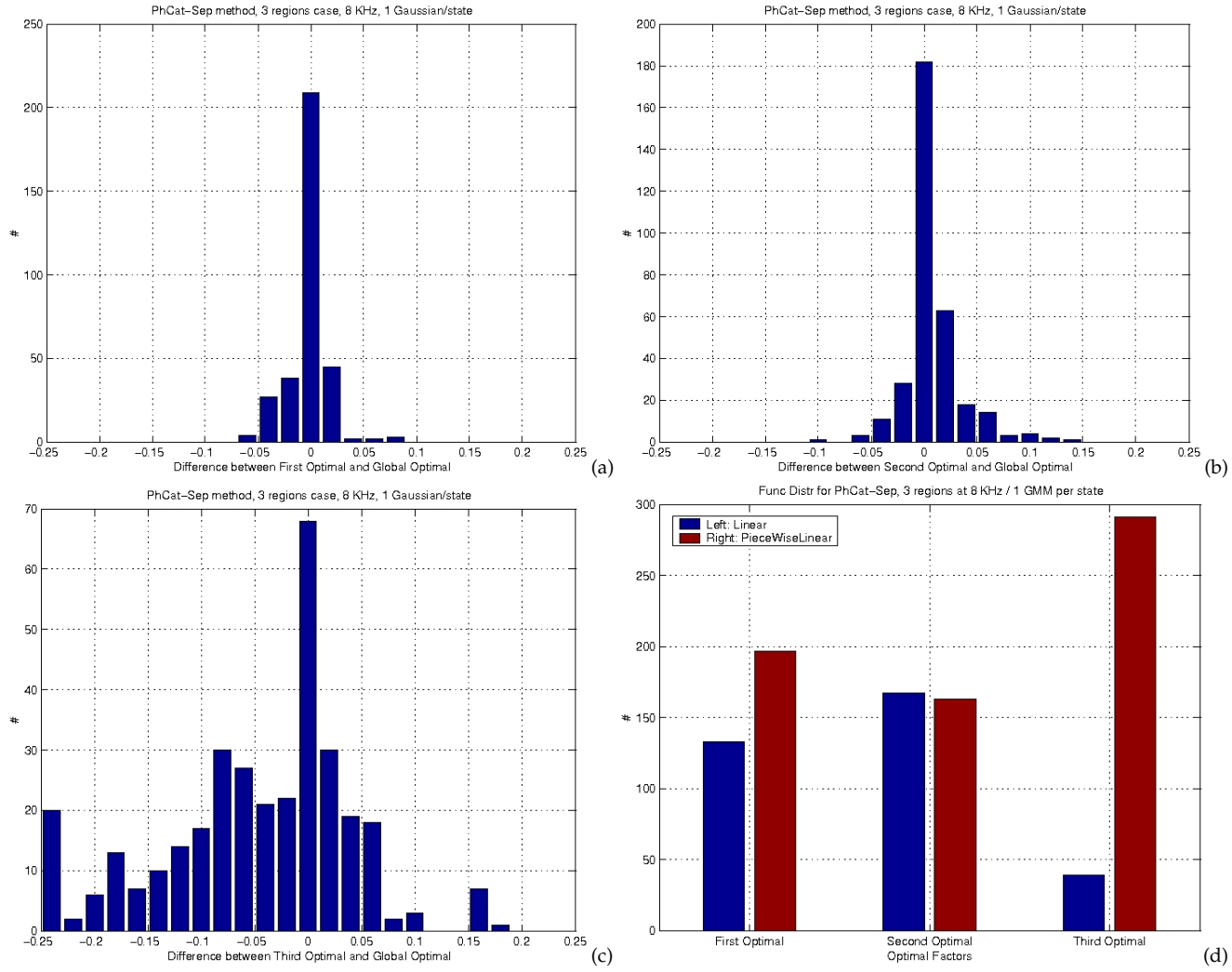


Figure 21. Optimal Factors Distribution For the Three Regions Case ((a), (b) and (c)) and the *PhCat-Sep* method. Also, the distribution of the optimal warping functions for each region (d). The sampling frequency is equal to 8 kHz.

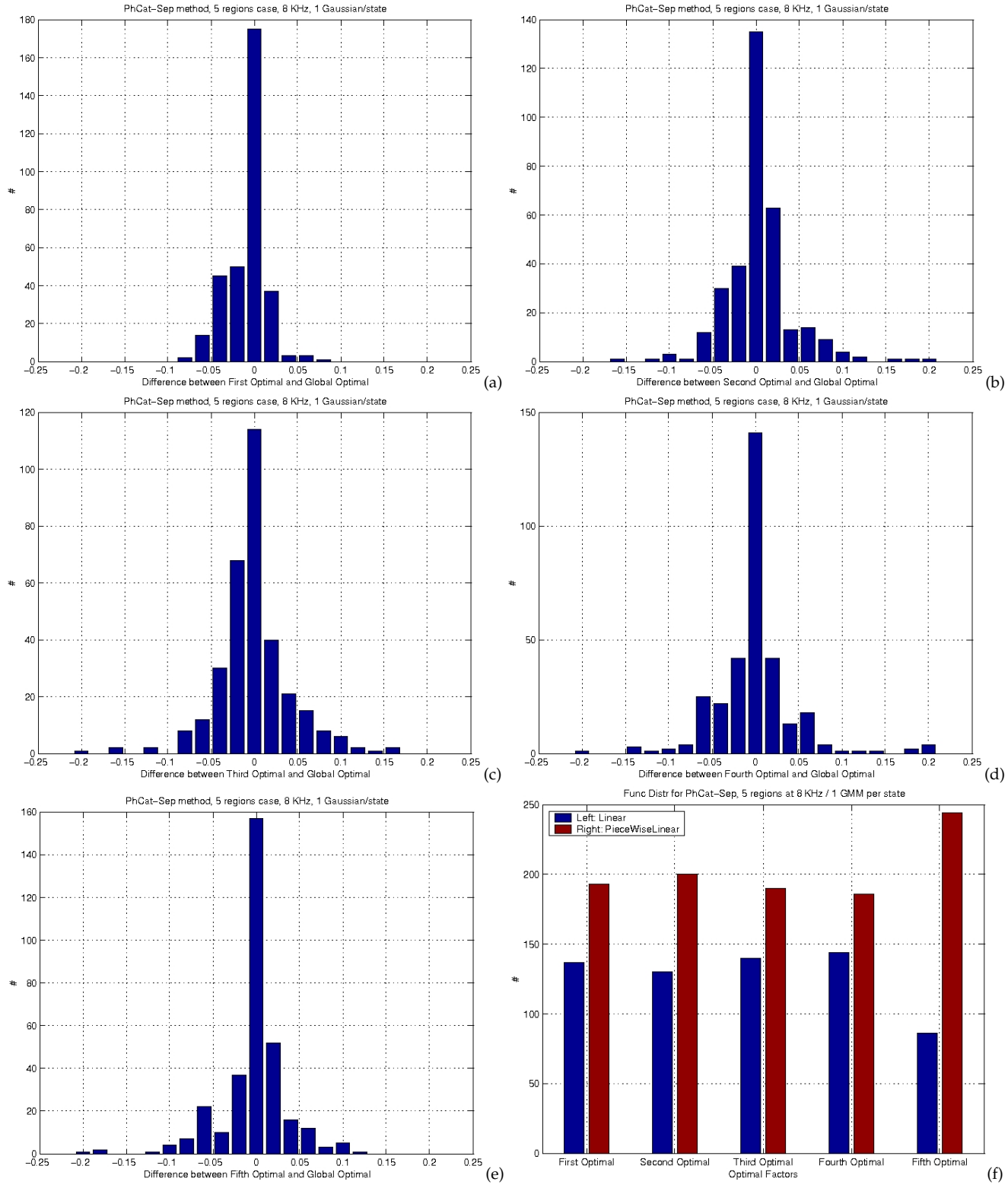


Figure 22. Optimal Factors Distribution For the Five Regions Case and the *PhCat-Sep* method. The sampling frequency is equal to 8 kHz.

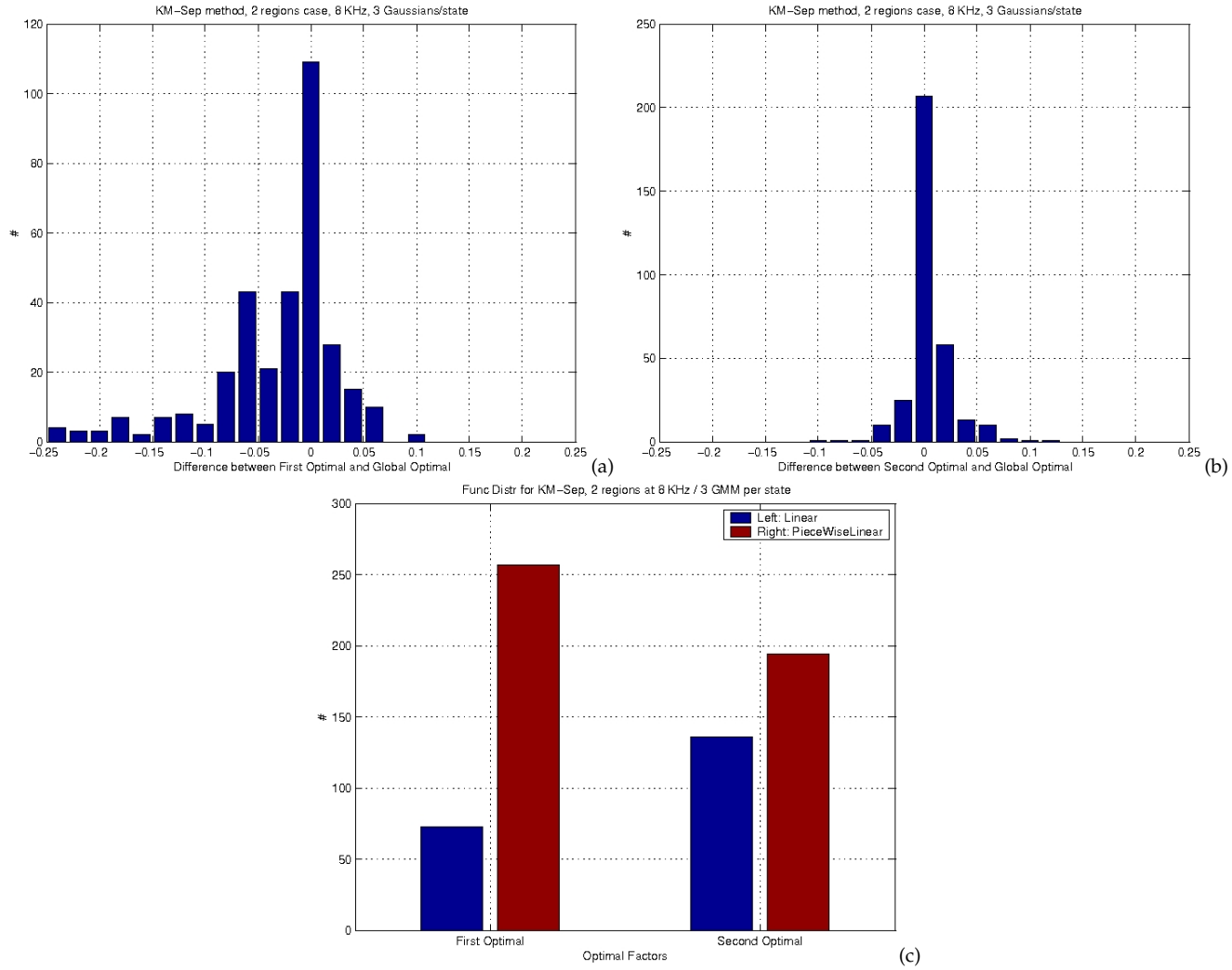


Figure 23. Distribution of the Difference Between the First Optimal Factor (a), the Second (b) with the Global Factor α_{glb} For the Two Regions Case and the *KM-Sep* method when the Gaussian Mixtures are equal to three. Also, the distribution of the chosen as optimal warping function for the two regions (c). The sampling frequency is equal to 8 kHz.

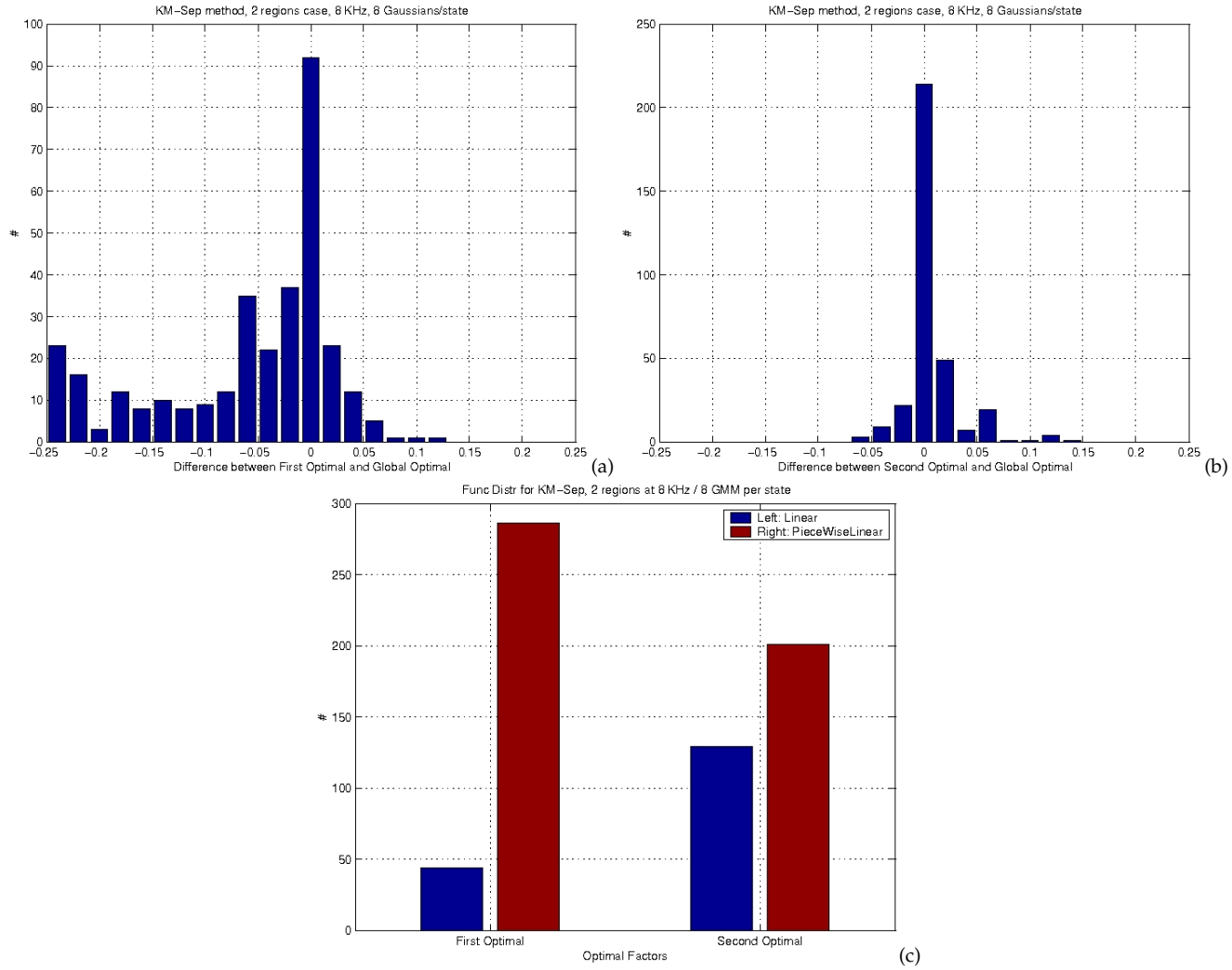


Figure 24. Distribution of the Difference Between the First Optimal Factor (a), the Second (b) with the Global Factor α_{glb} For the Two Regions Case and the *KM-Sep* method when the Gaussian Mixtures are equal to eight. Also, the distribution of the chosen as optimal warping function for the two regions (c). The sampling frequency is equal to 8 kHz.

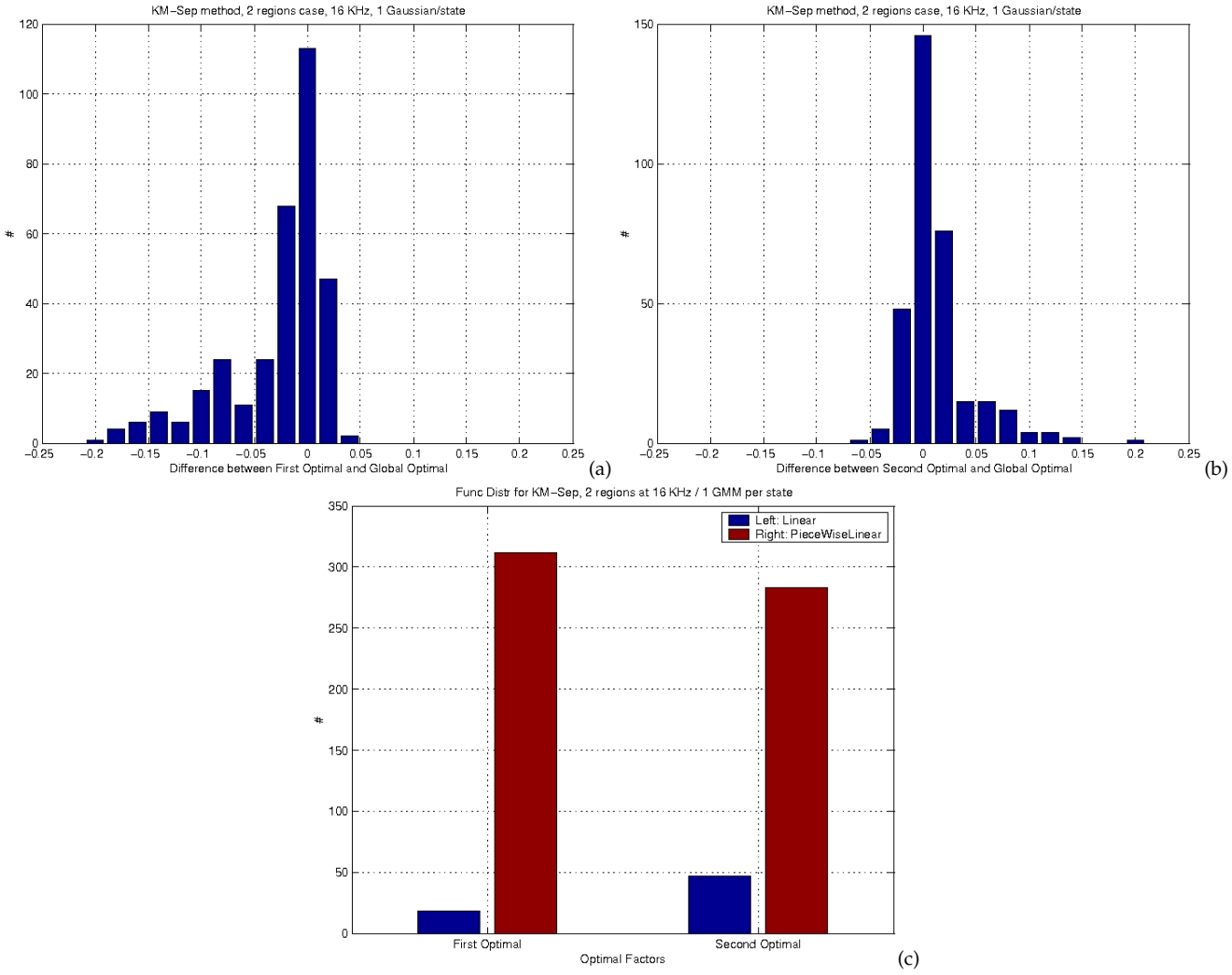


Figure 25. Distribution of the Difference Between the First Optimal Factor (a), the Second (b) with the Global Factor α_{glb} For the Two Regions Case and the *KM-Sep* method. Also, the distribution of the chosen as optimal warping function for the two regions (c).

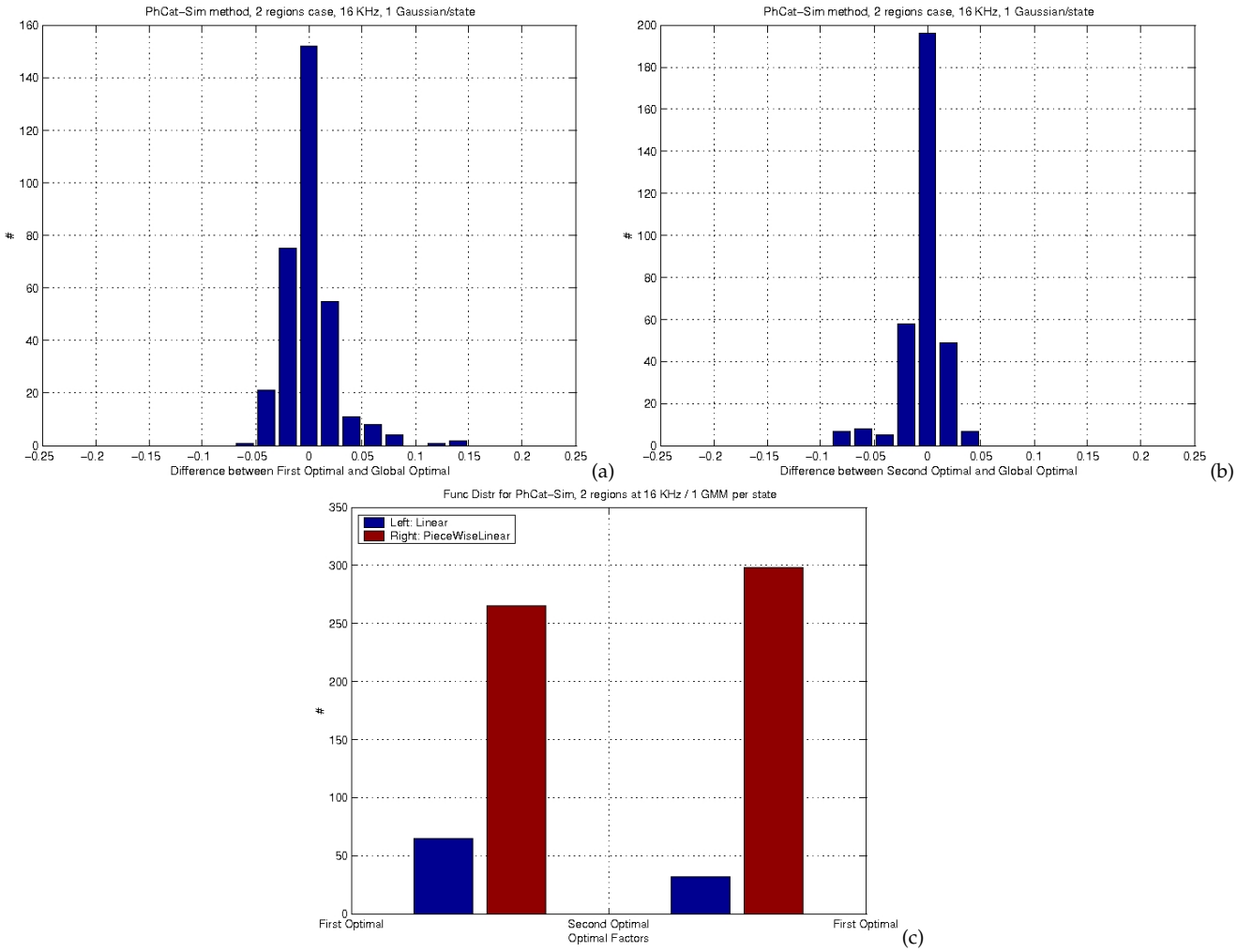


Figure 26. Distribution of the Difference Between the First Optimal Factor (a), the Second (b) with the Global Factor α_{glb} For the Two Regions Case and the *PhCat-Sim* method. Also, the distribution of the chosen as optimal warping function for the two regions (c).

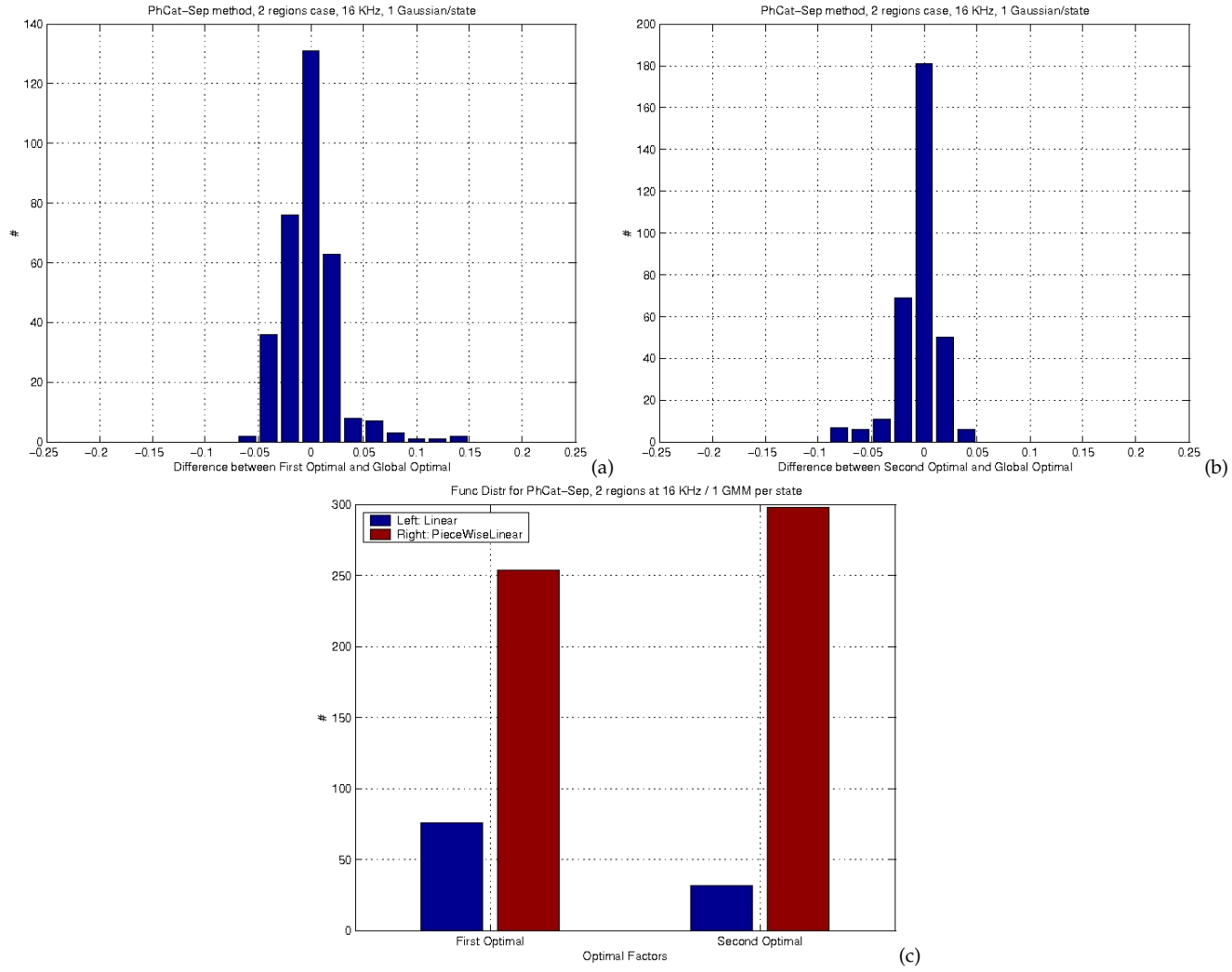


Figure 27. Distribution of the Difference Between the First Optimal Factor (a), the Second (b) with the Global Factor α_{glb} For the Two Regions Case and the *PhCat-Sep* method. Also, the distribution of the chosen as optimal warping function for the two regions (c).

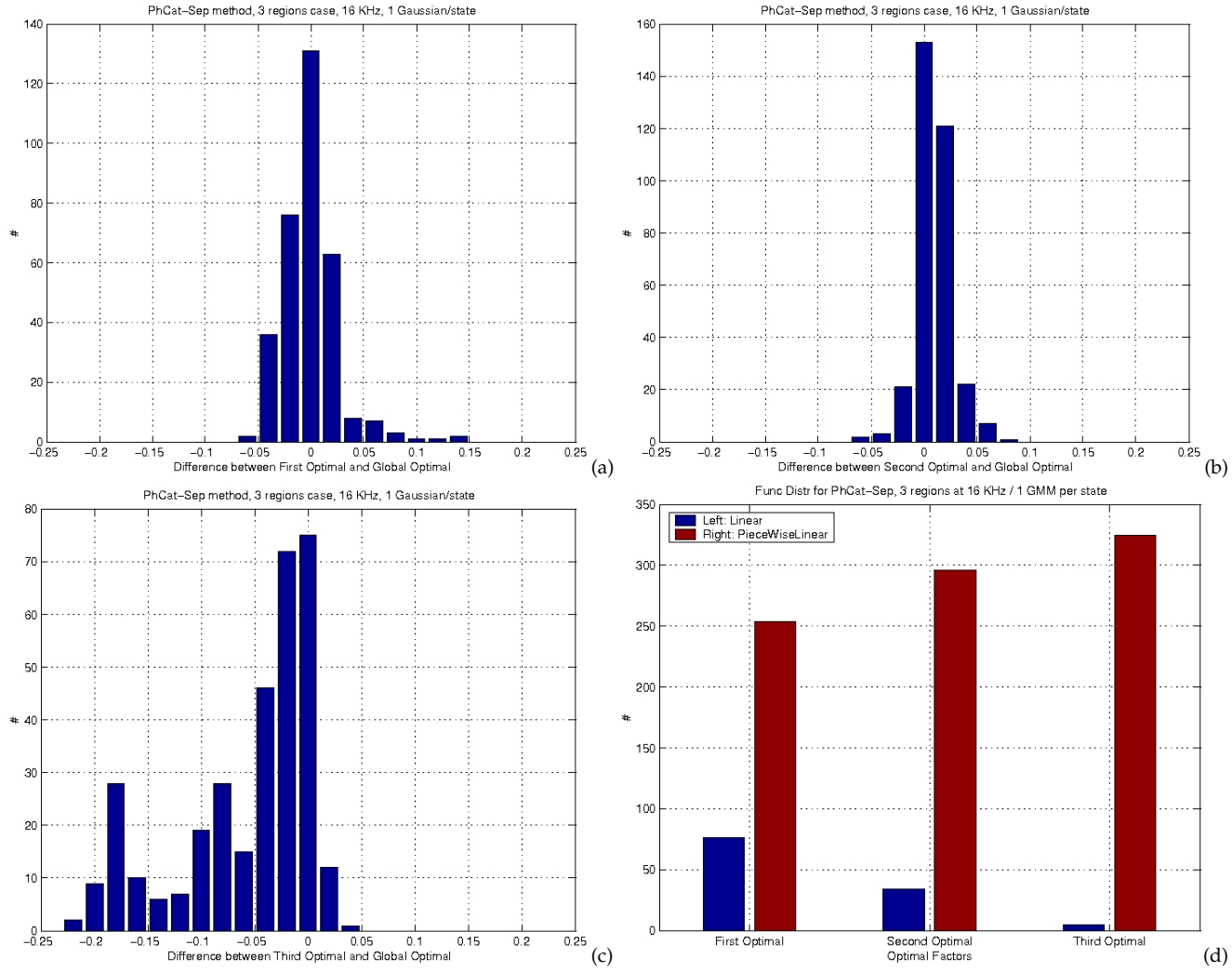


Figure 28. Optimal Factors Distribution For the Three Regions Case ((a), (b) and (c)) and the *PhCat-Sep* method. Also, the distribution of the optimal warping functions for each region (d). The sampling frequency is equal to 16 kHz.

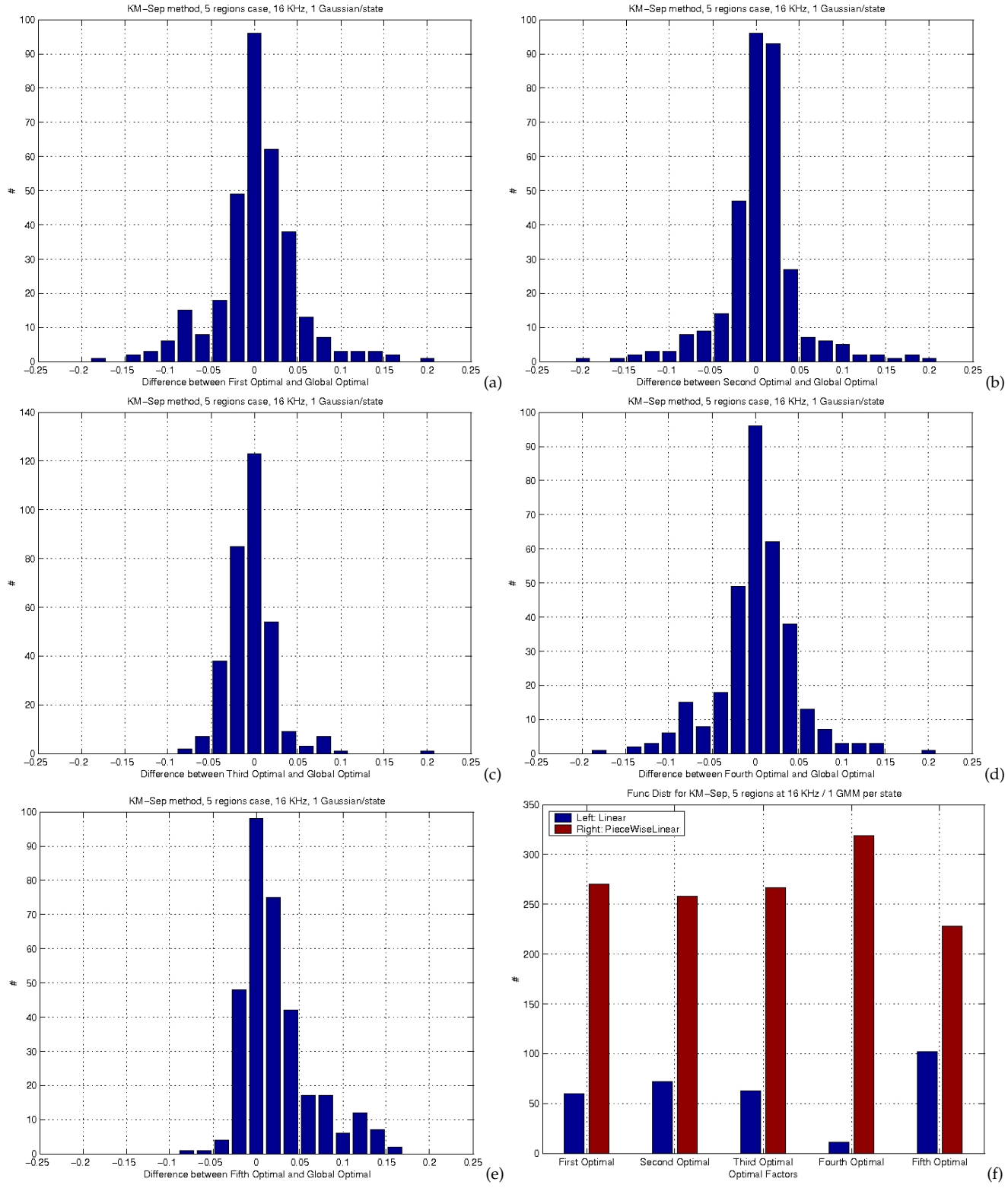


Figure 29. Optimal Factors Distribution For the Five Regions Case and the *PhCat-Sep* method.

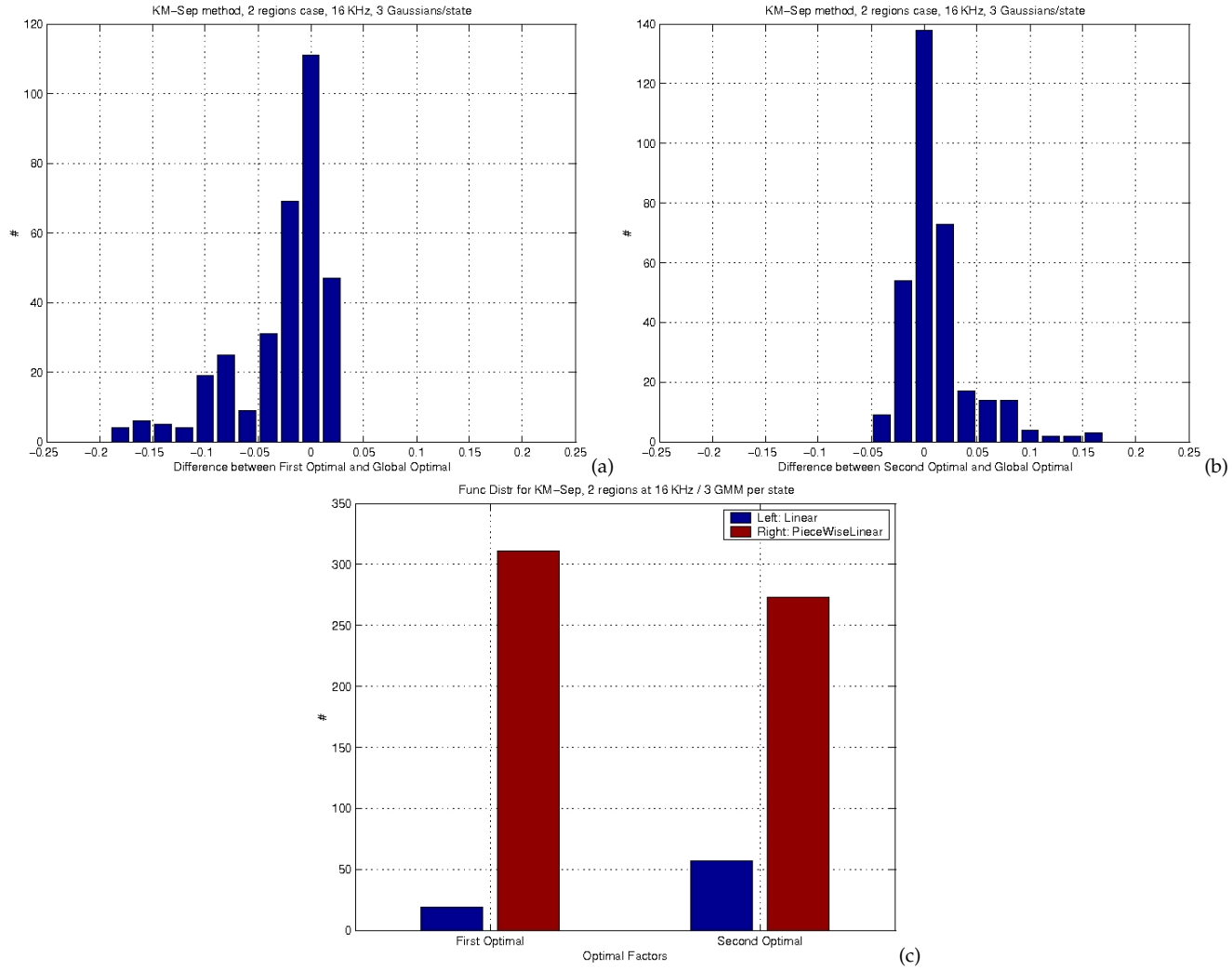


Figure 30. Distribution of the Difference Between the First Optimal Factor (a), the Second (b) with the Global Factor α_{glb} For the Two Regions Case and the *KM-Sep* method when the Gaussian Mixtures are equal to three. Also, the distribution of the chosen as optimal warping function for the two regions (c). The sampling frequency is equal to 16 kHz.

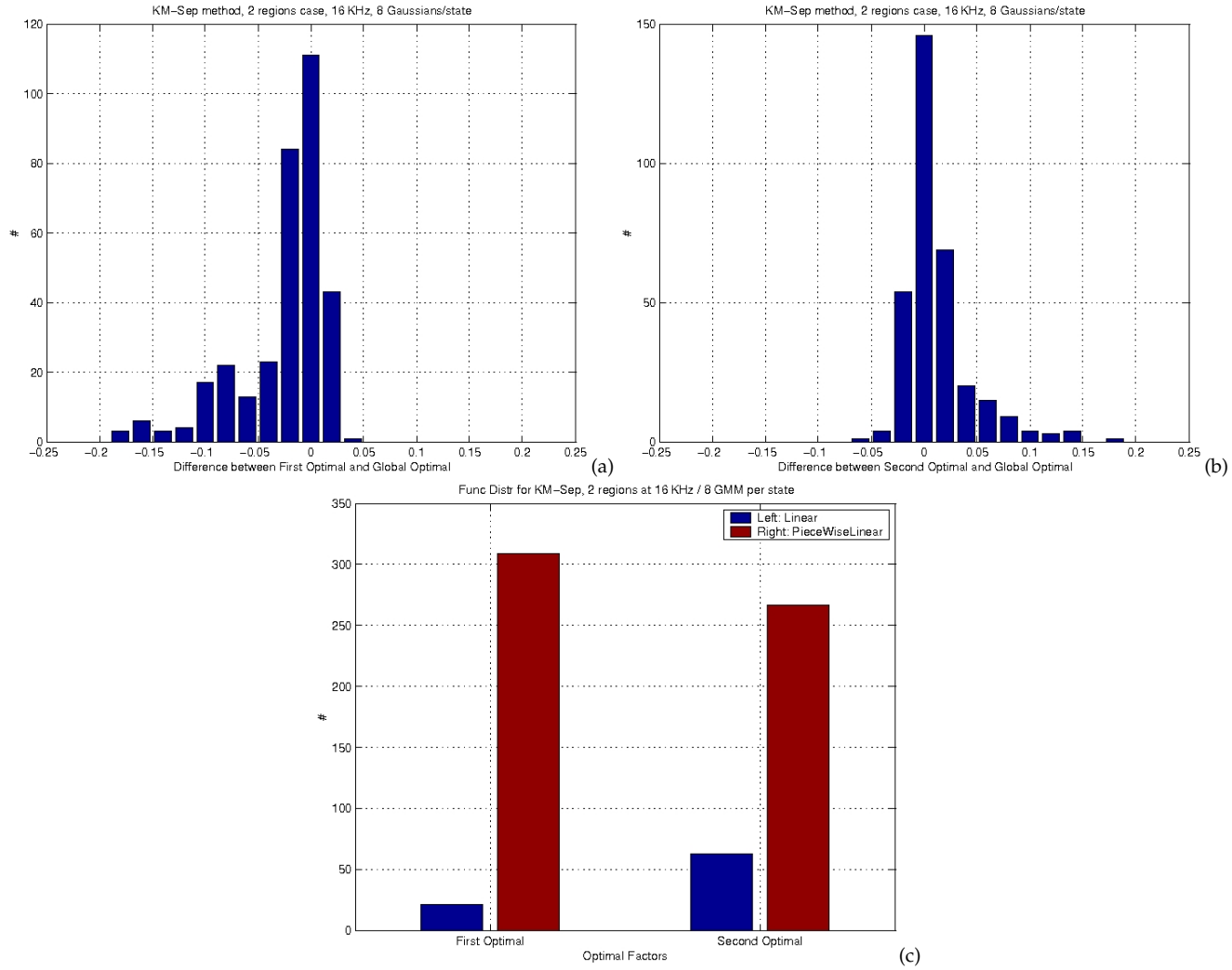


Figure 31. Distribution of the Difference Between the First Optimal Factor (a), the Second (b) with the Global Factor α_{glb} for the Two Regions Case and the *KM-Sep* method when the Gaussian Mixtures are equal to eight. Also, the distribution of the chosen as optimal warping function for the two regions (c). The sampling frequency is equal to 16 kHz.

9 DATABASES

9.1 TIMIT

TIMIT contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. A speaker's dialect region is the geographical area of the U.S. where they lived during their childhood years. The geographical areas correspond with recognized dialect regions in U.S. (Language Files, Ohio State University Linguistics Dept., 1982), with the exception of the Western region (dr7) in which dialect boundaries are not known with any confidence and dialect region 8 where the speakers moved around a lot during their childhood. Like TIDigits, it was recorded at TI; most of the labeling was carried out at the Massachusetts Institute of Technology (MIT). In contrast to the other corpora, the percentage of female speakers is only 30%.

Table 9 shows the chosen for our experiments speakers's information used in DARPA TIMIT Acoustic-Phonetic Speech Corpus. The initials denote:

- ID - Speaker initials (of form AAAN where A is alphabetic initial and N is a digit 0-9 to disambiguate identical initials
- Sex - Speaker gender (M or F)
- DR - Speaker dialect region number
    1. New England
    2. Northern
    3. North Midland
    4. South Midland
    5. Southern
    6. New York City
    7. Western
    8. Army Brat (moved around)

  At the final column (*Usage*), we present which of the speakers were used at our experiments as "reference" speakers and which of the speakers were used as "mapped".

We chose for "reference" and "mapped" speakers the following one:

The TIMIT corpus includes several files associated with each utterance. In addition to the speech waveform file (.wav) sampled at 16 KHz, three associated transcription files (.txt, .wrd, .phn) exist. These associated files have the form:

SPEAKERS			
ID	Sex	DR	Usage
cjfo	F	1	Reference
dxwo	F	2	Reference
grwo	F	3	Reference
klco	F	4	Reference
bbro	M	7	Reference
hito	M	5	Reference
sdso	M	6	Reference
tcso	M	8	Reference
dawo	F	1	Mapping
scno	F	2	Mapping
ltmo	F	3	Mapping
pafo	F	4	Mapping
dhlo	M	5	Mapping
keso	M	6	Mapping
bomo	M	7	Mapping
rreo	M	8	Mapping

Table 9. Reference and Mapped Speakers.

- .wav : SPHERE-headered speech waveform file sampled at 16 kHz.
- .txt : Associated orthographic transcription of the words the person said. (Usually this is the same as the prompt, but in a few cases the orthography and prompt disagree.)
- .wrd : Time-aligned word transcription. The word boundaries were aligned with the phonetic segments using a dynamic string alignment program.
- .phn : Time-aligned phonetic transcription.

9.2 AURORA4 DATABASE

The WSJo (Wall Street Journal) database (available from LDC under the name CSR-I (WSJo) complete) is chosen as basis for the experiments. The recognition of a 5000 word vocabulary is selected as task as it has been also used for the ARPA evaluations on continuous speech recognition. Besides the original data sampled

at 16 kHz a second version of the data is created by downsampling the data from 16 kHz to 8 kHz. The reason for this is because most of today's telecommunication terminals operate in the frequency range up to 4 kHz. But future speech services will aim at a higher speech intelligibility and a higher subjective speech quality by analysing speech at a higher bandwidth up to 8 kHz.

The WSJ data have been recorded with a Sennheiser microphone and with a second microphone in parallel. The recordings with the second microphone are used for enabling recognition experiments with different frequency characteristics in the transmission channel.

We define training mode which takes clean data only to train the recognizer. The predefined ARPA test set is selected here to perform the recognition on a 5000 word vocabulary.

9.2.1 *Filtering.*

An additional filtering is applied to consider the realistic frequency characteristics of terminals and equipment in the telecommunication area. Two "standard" frequency characteristics are used which have been defined by the ITU. The abbreviations G.712 and P.341 have been introduced as reference to these filters. The G.712 characteristic is defined for the frequency range of the usual telephone bandwidth up to 4 kHz and has a flat characteristic in the range between 300 and 3400 Hz. P.341 is defined for the frequency range up to 8kHz and represents a band pass filter with a very low cut off frequency at the lower end and a cut off frequency at about 7 kHz at the higher end of the bandpass.

9.2.2 *Training and Testing sets.*

The WSJ0 database (AURORA4 is part of extended database WSJ0) consists of speech that has been recorded with two microphones in parallel. The first one is a close- talking microphone of type Sennheiser HMD414. No regulation exists about the choice of the second microphone which has been e.g. a desk mounted microphone. Most recordings consist of read texts from the Wall Street Journal.

In the ARPA evaluations a set of about 7200 utterances (about 12 hours of speech) has been selected for the training. 7138 recordings are available. These data are taken from the recordings with the Sennheiser microphone. We consider the same set for our training mode on clean data only. We refer to this training mode by naming it "training_clean_sennh".

A set of 330 utterances has been designated in the ARPA evaluation to perform a baseline recognition on the 5000 word vocabulary. This test includes the usage of a closed vocabulary bigram language model as supplied by Lincoln. The 330 utterances contain recordings from 8 speakers with about 40 utterances per speaker. The test set that we used for evaluation is the one which contained clean filtered data at 8 KHz with Sennheiser microphone.

Testing Speaker information are followed at Table 10. The keys which identify the presented information are:

- ID : speaker ID.
- Gender : speaker gender, Male (M) or Female (F).

SPEAKERS	
ID	Gender
440	M
441	F
442	M
443	M
444	F
445	F
446	M
447	M

Table 10. Information (Gender) for Testing Speakers at AURORA4.

BIBLIOGRAPHY

[1] Alexandros Potamianos, Richard Rose. On Combining Frequency Warping and Spectral Shaping in HMM Based Speech Recognition. *Proc. Internat. Conf. on Acoust., Speech, and Signal Process, Munich, 1997*. (Cited on page 11.)

[2] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, Phil Woodlan. The HTK Book (for HTK Version 3.3). *Cambridge University Eng. Deptï¿½Speech Group*. (Cited on page 3.)

[3] L. Rabiner, B.-H. Juang. Fundamentals of Speech Recognition. *Prentice-Hall, NJ, USA, 1993*. (Cited on page 4.)

[4] Pierre Dognin, A BandPass Transform for Speaker Normalization *PHD Thesis, University of Pittburgh, 2003*. (Cited on page 9.)

[5] Shizhen Wang, Xiaodong Cui, Abeer Alwan, Speaker Adaptation With Limited Data Using Regression-Tree-Based Spectral Peak Alignment *INTERSPEECH, 2006*. (Cited on page 49.)

[6] Michael Pitz, Sirko Molau, Ralf SchlÃ¼ter, Hermann Ney, Vocal Tract Normalization Equals Linear Transformation in Cepstral Space *in Proc. of the EUROSPEECH 2001, Aalborg, Denmark, 2001*. (Cited on page 13.)

[7] Tadashi Emori, Koichi Shinoda newblock Rapid Vocal Tract Length Normalization using Maximum Likelihood Estimation *in Proc. of the EUROSPEECH 2001, Aalborg, Denmark, 2001*. (Cited on page 13.)

[8] Davis, S.B., Mermelstein, P., Comparison of Parametric Repre-sentations for Monosyllabic Word Recognition in Continuously Spoken Sentences IEEE Trans. on Acoustic, Speech and Signal Processing, 1980. (Cited on page 5.)

[9] E. Eide and H. Gish. A parametric approach to vocal tract length normalization. *Proc. ICASSP*, 8(2):346–348, May 1996. (Cited on pages 9, 11, and 12.)

[10] Ev. B. Gouvêa and R. M. Stern. Speaker normalization through formant-based warping of the frequency scale. *Proc. Eurospeech '97, Rhodes, Greece*, 8(2), May 1997. (Cited on pages 10 and 12.)

[11] L. Lee and R. Rose. A frequency warping approach to speaker normalization. *Proc. IEEE Trans. Speech and Audio Processing*, 6(2):49–60, January 1998. (Cited on page 9.)

[12] L. Lee and R. C. Rose. Speaker normalization using efficient frequency warping procedures. *Proc. ICASSP*, 8(2):353–356, May 1996. (Cited on pages 9, 10, 11, 28, 35, 36, and 47.)

[13] A. Miguel, E. Lieida, R. Rose, L. Buera, and A. Ortega. Augmented state space acoustic decoding for modeling local variability in speech. *Proc. INTERSPEECH*, pages 3009–3012, 2005. (Cited on page 14.)

[14] S. Molau, St. Kanthak, and H. Ney. Efficient vocal tract normalization in automatic speech recognition. *Proc. EUROSPEECH*, pages 2527–2530, March 1999. (Cited on page 19.)

[15] S. Panchapagesan and Ab. Alwan. Multi-parameter frequency warping for vtln by gradient search. *Proc. ICASSP*, pages 1181–1184, 2006. (Cited on pages 10 and 15.)

[16] L.R. Rabiner and R.W. Schafer. *Digital Processing of Speech Signals*. Prentice Hall,, Upper Saddle River, New Jersey, 4th edition, 1978. (Cited on page 3.)

[17] L.R. Rabiner and R.W. Schafer. *Digital Processing of Speech Signals*. Prentice Hall,, Upper Saddle River, New Jersey, 4th edition, 1978.

[18] J. Orloff S. Wegmann, D. McAllaster and B. Peskin. Speaker normalization on conversational telephone speech. *Proc. ICASSP*, pages 339–341, May 1996. (Cited on pages 9 and 12.)

[19] Ok Keun Shin. A vector-quantizer based method of speaker normalization. *Proc. ICIS*, pages 402–407, 2005. (Cited on page 14.)

[20] H. Wakita. Normalization of vowels by vocal tract length and its application to vowel identification. *Proc. ICASSP*, ASSP–25:183–192, April 1977. (Cited on pages 9 and 35.)

[21] L. Welling, S. Kanthak, and H. Ney. Improved methods for vocal tract normalization. *Proc. ICASSP*, pages 761–764, March 1999. (Cited on pages 10, 11, and 12.)

[22] P. Zhan and A. Waibel. Vocal tract length normalization for large vocabulary continuous speech recognition. *Proc. ICASSP*, 8(2), May 1997. (Cited on page 10.)

[23] P. Zhan and M. Westphal Speaker Normalization Based on Frequency Warping. *Proc. ICASSP*, May 1997. (Cited on pages 12 and 18.)

[24] Westphal M., Schultz T., Waibel A. Linear Discriminant. A New Criterion For Speaker Normalization *Proc. ICSLP '98, pp. 827-830, Australia* (Cited on page 13.)

[25] Hans Dolfing Exhaustive Search for Lower-Bound Error-Rates in Vocal Tract Length Normalization *In ICSLP-2000, vol.1, 762-765.* (Cited on page 13.)

[26] Alexandros Potamianos, Shrikanth Narayanan Robust Recognition Of Children's Speech *Proc. ICASSP, 2003.* (Cited on page 16.)

[27] L. Welling, H. Ney, S. Kanthak Speaker Adaptive Modeling by Vocal Tract Normalization *IEEE Trans on Speech and Audio Processing, vol. 10, pp. 415-426, Sep. 2002* (Cited on page 12.)

[28] Stephen Cox Speaker Normalization in the MFCC Domain *In ICSLP-2000, vol.3, 853-856.* (Cited on page 13.)

[29] L. Gillick, Stephen Cox, Some Statistical Issues In The Comparison of Speech Recognition Algorithms *ICASSP, 1989.* (Cited on page 47.)

[30] Davis and Mermelstein, Comparison of Parametric Representations for Monosyllable Word Recognition in Continuously Spoken Sentences *IEEE Transactions on Acoustic, Speech and Signal Processing, 1980.* (Cited on page 4.)

[31] Jonas Loof, Hermann Ney, Srinivasan Umesh Vtln Warping Factor Estimation Using Accumulation Of Sufficient Statistics *ICASSP 2006* (Cited on page 15.)

[32] S. V. Bharath Kumar, S. Umesh, R. Sinha Study Of Non-Linear Frequency Warping Functions For Speaker Normalization *ICASSP 2006* (Cited on page 15.)

[33] L. Lee I'm sorry Dave, I'm afraid I can't do that: Linguistics, Statistics and Natural Language Processing circa 2001 *Computer Science: Reflections on the Field, Reflections from the Field* (Cited on page 8.)

[34] F.S. Chen, Goodman,. J. An Empirical Study of Smoothing Techniques for Language Modeling *In: Proc. Thirty-Fourth Annual Meeting of the Association for Computational Linguistics* (Cited on page 8.)

[35] Rosenfeld. R. Two Decades of Statistical Language Modeling: where do we go from here? *Proceedings of the IEEE, Vol.88, Iss.8* (Cited on page 7.)

[36] Papoulis. A., Pillai, U.S. Probability, Random Variables and Stohastic Processes *McGraw-Hill* (Cited on page 8.)

[37] Sankaran Panchapagesan Frequency warping by linear transformation of standard MFCC *INTERSPEECH-2006*