# Local Transformation Models for Speech Recognition

*Antonio Miguel, Eduardo Lleida, Alfons Juan*, Luis Buera, Alfonso Ortega and Oscar Saz*

I3A, University of Zaragoza, Spain
{amiguel,lleida,lbuera,ortega,oskarsaz}@unizar.es

*DSIC, Polytechnic University of Valencia, Spain
ajuan@dsic.upv.es

## Abstract

This paper presents a novel acoustic modeling framework that naturally extends the Hidden Markov Model (HMM) approach. The novel models reduce the errors caused by speaker variability by means of a local spectral mismatch reduction. A more complex and flexible speech production scheme can be assumed, in which the local temporal and frequency elastic deformations of the speech are captured by the model. In the new framework the states of a standard HMM, which are usually associated with temporal transitions, are expanded so that a new degree of freedom for the model is provided and it is then possible to estimate an optimum frequency warping factor at the same time as the decoder finds the best state sequence. In the local spectral warping based models the states become time-frequency related states and the number of parameters of the model is comparable to the standard HMM since they share a certain amount of parameters as it will be shown. The novel models are evaluated in the noise-free TIDIGITS corpus, which includes connected digits uttered by male, female and children. It has been found that, under speaker group (age-gender) mismatch conditions, the local frequency warping reduced Word Error Rate (WER) in mean by a $70\%$, using the initial models. When matched speaker group conditions were tested the error was reduced in mean in a $9.7\%$ after reestimating the models.
**Index Terms**: speaker variability, local frequency warping.

## 1. Introduction

A speech modeling technique for speaker variability reduction is investigated in this paper, since this variability has a great interest due to the impact on the accuracy of Automatic Speech Recognition, ASR, systems. It will be shown that the model presented provides a mechanism to reduce the error on ASR for a wide range of local deformations of the speech parameters across the time and frequency axes.

Standard techniques as Hidden Markov Models (HMM) provide a successful reduction of the speaker variability in terms of temporal variability thanks to the time alignment of the utterances to the models by the Viterbi algorithm, capturing the essential information needed for speech recognition tasks. In the HMM framework there also exists a basic mechanism to model the frequency variability due to speaker, which causes changes in the vocal tract shape. It is provided by the state dependent observation generating process, which usually is assumed to follow a probability density function pdf as a Gaussian Mixture Model (GMM) The vocal tract shape deviations due to a large population of speakers are captured by the state pdfs as different components of the mixture. Then, a number of examples from each one of the shapes are needed so that the components of the mixture can be estimated in the learning process. Therefore a large amount of Gaussian components and training data are required in order to deal with this source of variability in a simple HMM.

Some methods have appeared in order to compensate more accurately both sources of variability specially in the frequency axis. In this paper we focus the experiments towards this kind of speaker variability, manifested as the frequency deformations of the spectrum envelope that occur in speaker independent ASR tasks, which are known to have its origin in the vocal tract and articulatory instant shapes. Some methods have appeared previously in order to compensate for speaker frequency variability as Vocal Tract Length Normalization, VTLN, [1, 2] and Maximum Likelihood Linear Regression, MLLR, [3], which reduce the mismatch between data an model, but those methods compensate the mismatch given previous utterances and transcriptions or extra speaker dependent training data. The model framework, referred from here as the augMented stAte space acousTic dEcoder/modEl (MATE), consists of an expansion of the VTLN methods to provide local transformations to be locally optimized, simultaneously to the decoding of the state sequence in an expanded search trellis. The training and the testing of MATE is speaker independent, since it is expected to capture part of the speaker variability by means of the expanded state space and the inter-transformation transitions.

The first approaches to this paradigm were envisioned in [4] and then followed by [5, 6] in a more general approach. Those methods were intended to normalize the speech signal to be better accepted by the model. The model presented in this paper is an evolution of them and the transformation is embedded into the model, allowing a more general formulation and derivation of the model parameter estimation expressions, as it will be shown in sections 3 and 3.2. The transformations of MATE described in this article are a valid generalization of [5] in both sources of variability, time and frequency but, as the effect of the local temporal warping is less noticeable unless a stressed or pathological speech corpus is tested, the experiments in this article are going to be oriented to show speaker independent ASR improvements in the sense of frequency transformations in the new MATE framework.

The paper is organized as follows. Section 2 reviews the the existing techniques used for speaker mismatch reduction. Section 3 presents the model formulation and the procedure for estimating the model parameters using the EM algorithm. Section 4 includes the results of an experimental study of the new models. Finally, discussion and conclusions are presented in Section 5.

## 2. Speaker mismatch reduction model based methods

Basic HMM provides a simple, but effective under certain conditions, mechanism of modeling speaker variability which consists

This work has been supported by the national project TIN 2005-08660-C04-01

of learning the observations generated by the same state in the model thanks to a multimodal pdf, as the mixture of Gaussians. But in order to reduce speaker variability mismatch in a more general way, frequency warping based speaker normalization techniques as VTLN, have been applied in many ASR task domains [1]. This class of techniques produces a warped frequency scale by selecting an optimum warping function chosen to maximize the average likelihood of the normalized sequence with respect to the HMM.

In [7], it was shown that the procedure for obtaining the frequency warped features was equivalent to a linear projection of the original cepstrum. This work revealed a straight forward relationship between VTLN and MLLR methods.

Both methods reduce the need of large amounts of data to train speaker independent models but suffer from two main problems. The first one is that it is generally implemented as a two pass procedure which can make real-time implementation difficult. The first pass is used to generate an initial hypothesized word string. This initial word string is then used in a second pass to normalize the data or the models to reduce the mismatch. The second limitation is related to the fact that only a single linear warping function or model transformation is selected for the whole utterance. Even though physiological evidence indicates that all phonetic events do not exhibit similar spectral variation as a result of physiological differences on vocal tract shape. The procedure described in [5] and generalized in this article, addressed both of these issues and it showed a good performance when compared to previous methods. The procedure requires only a single pass over the input utterance and produces frame-specific estimates of the frequency warping functions.

# 3. MATE

In order to model the vocal tract shape changes during speech utterances and across speakers we propose a model (MATE) in which a new degree of freedom has been added to track those changes in a HMM way.

Following [7], the spectral warping performed in the previous decoding method [5] can be seen as a linear projection of the cepstral feature space, $\mathbf{X}^{\alpha_n} = \mathbf{A}_n\mathbf{X}$, with $n = 1, \cdots, N$, the number of warping factors.

The model is constructed after a state space expansion that is similar to [5], where a state $q$ is expanded into states $(q, n)$ with $n$ the index of the transformations. The new model provides observation generation pdfs in the states that depend on a discrete set of transformation matrices, $\{\mathbf{A}_n\}_{n=1}^{N}$, embedding the warping in the model as a general transformation instead of normalizing data as before [5].

Given that a component in the original state pdf mixture follows normal distribution: $\mathcal{N}(\mu_q, \mathbf{\Sigma}_q)$, the expanded states components are assumed to follow a distribution:

$$\mathbf{x}_t|_{n,q} \sim \mathcal{N}(\mathbf{A}_n\mu_q, \mathbf{A}_n\mathbf{\Sigma}_q\mathbf{A}_n^t), \tag{1}$$

so that the model can generate sequences of warped cepstrum vectors, which we expect to be closer to real data.

### 3.1. Complete model

For clarity in the hidden variable derivation lets firstly assume that a complete set of labeled data is available, the joint pdf of the data and label sequences is called complete or visible model. The sequences that could be generated by such model are: a cepstrum data sequence, $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_T)^t$, a state labels sequence, $\mathbf{S} = (\mathbf{s}_1, ..., \mathbf{s}_T)^t$, the transformation labels sequence,$\mathbf{R} = (\mathbf{r}_1, ..., \mathbf{r}_T)^t$, where $\mathbf{x}_t \in \mathbb{R}^D$ (with $D$ the dimension of the feature vector), $\mathbf{s}_t$ is an indicator vector $\mathbf{s}_t \in \{0, 1\}^Q$, with 1 in the state index $q$ that generated the observation $\mathbf{x}_t$ and zeros elsewhere as in [8], and finally $\mathbf{r}_t$ is another indicator vector $\mathbf{r}_t \in \{0, 1\}^N$, with 1 in the transformation matrix index $n$ used to generate the observation $\mathbf{x}_t$ and zeros elsewhere.

The pdf of a sequence of this kind can be written as follows using the rule of Bayes,

$$\begin{aligned} f(\mathbf{X}, \mathbf{S}, \mathbf{R}) &= f(\mathbf{S}, \mathbf{R}) f(\mathbf{X}|\mathbf{S}, \mathbf{R}) \\ &= \prod_{t\geq 1}^{T} f(\mathbf{s}_t, \mathbf{r}_t|\mathbf{s}_1^{t-1}, \mathbf{r}_1^{t-1}) \prod_{t\geq 1}^{T} f(\mathbf{x}_t|\mathbf{x}_1^{t-1}, \mathbf{S}, \mathbf{R}). \end{aligned} \tag{2}$$

Taking HMM-like assumptions, we can approximate (2) by:

$$f(\mathbf{X}, \mathbf{S}, \mathbf{R}) = \prod_{t\geq 1}^{T} f(\mathbf{s}_t, \mathbf{r}_t|\mathbf{s}_{t-1}, \mathbf{r}_{t-1}) \prod_{t\geq 1}^{T} f(\mathbf{x}_t|\mathbf{s}_t, \mathbf{r}_t). \tag{3}$$

The indicator vectors follow a Multinomial distribution of parameters,

$$\mathbf{\Pi} = \{\pi_{q,n,q',n'}\}_{q=1,n=1,q'=1,n'=1}^{Q,N,Q,N}, \tag{4}$$

being $\pi_{q,n,q',n'}$ the transition from state $(q, n)$ to $(q', n')$ probability,

$$f(s_{t,q'} = 1, r_{t,n'} = 1|s_{t-1,q} = 1, r_{t,n} = 1) = \pi_{q,q',n,n'}. \tag{5}$$

Making use of it and taking into account that the indicator variables are zeros in all positions except one, then we can express (3) as (6), where the expanded state $(q, n)$ pdf in (6), $f(\mathbf{x}_t|s_{t,q} = 1, r_{t,n} = 1)$ , follows a distribution of the form of (1). The ensemble of parameters composed by $\mathbf{\Pi}$ and the state pdfs are referred as $\mathbf{\Theta}$.

### 3.2. EM training algorithm

When the labeled data of the complete problem is missing as in speech applications, $\mathbf{S}$ and $\mathbf{R}$ are hidden variables, the EM is a well known algorithm that provides a method for estimating the parameters of the model in an iterative two step process.

The first step, E expectation step, consists of calculating the auxiliary function $Q(\mathbf{\Theta}|\mathbf{\Theta}^{(k)}) = E[\log f(\mathbf{X}, \mathbf{S}, \mathbf{R}|\mathbf{\Theta})|\mathbf{X}, \mathbf{\Theta}^{(k)}]$ that involves expected value computations for the hidden variables with respect to the data and the model parameters at iteration $k$. It can be expressed as in (7) for our model, where the expressions noted as $(\cdot)^{(k)}$ refer to the expected values of the variable between the parentheses:

$$\begin{aligned} (s_{t,q}r_{t,n})^{(k)} &= E[s_{t,q}r_{t,n}|\mathbf{X}, \mathbf{\Theta}^{(k)}] \\ &= f(s_{t,q} = 1, r_{t,n} = 1|\mathbf{X}, \mathbf{\Theta}^{(k)}). \end{aligned} \tag{8}$$

$$\begin{aligned} (s_{t-1,q}r_{t-1,n}s_{t,q'}r_{t,n'})^{(k)} &= E[s_{t-1,q}r_{t-1,n}s_{t,q'}r_{t,n'}|\mathbf{X}, \mathbf{\Theta}^{(k)}] \\ &= f(s_{t-1,q} = 1, r_{t-1,n} = 1, s_{t,q'} = 1, r_{t,n'} = 1|\mathbf{X}, \mathbf{\Theta}^{(k)}). \end{aligned} \tag{9}$$

Those expressions are difficult to calculate directly but thanks to the expanded auxiliary functions $\alpha_{t,q,n}$, $\beta_{t,q,n}$, which can be calculated recursively, computations are reduced to an affordable level. Nevertheless, in order to speed up the method and having experimented almost identical results, the expected values in (8) and (9) can be approximated in hard decision way, (0 or 1), by the Viterbi decoding algorithm.

The second step, M maximization step, consists of maximizing the $Q(\mathbf{\Theta}|\mathbf{\Theta}^{(k)})$ function with respect to the model parameters

$$f(\mathbf{X},\mathbf{S},\mathbf{R}) = \prod_{q,n} \pi_{0,0,q,n}{}^{s_{1,q}r_{1,n}} \prod_{t\geq 2} \prod_{q,q',n,n'} \pi_{q,n,q',n'}^{s_{t-1,q}r_{t-1,n}s_{t,q'}r_{t,n'}} \prod_{t\geq 1} \prod_{q,n} f(\mathbf{x}_t|s_{t,q}=1,r_{t,n}=1)^{s_{t,q}r_{t,n}}. \quad (6)$$

$$Q(\mathbf{\Theta}|\mathbf{\Theta}^{(k)}) = \sum_{q,n} (s_{1,q}r_{1,n})^{(k)} \log \pi_{0,0,q,n} + \sum_{t\geq 2} \sum_{q,q',n,n'} (s_{t-1,q}r_{t-1,n}s_{t,q'}r_{t,n'})^{(k)} \log \pi_{q,n,q',n'} + \sum_{t\geq 1} \sum_{q,n} (s_{t,q}r_{t,n})^{(k)} \log f(\mathbf{x}_t|s_{t,q}=1,r_{t,n}=1). \quad (7)$$

from each iteration in order to obtain the values for the parameters in the next iteration, $\mathbf{\Theta}^{(k+1)} = \arg\max_{\mathbf{\Theta}} Q(\mathbf{\Theta}|\mathbf{\Theta}^{(k)})$ Finally maximizing the expression subject to the constraint,

$$\sum_{q',n'} \pi_{q,n,q',n'} = 1, \forall q,n, \quad (10)$$

we obtain the following expressions for the parameter estimations in the iteration $k+1$, (for a single Gaussian model for simplicity):

$$\pi_{q,n,q',n'}^{(k+1)} = \frac{\sum_{t\geq 2}(s_{t-1,q}r_{t-1,n}s_{t,q'}r_{t,n'})^{(k)}}{\sum_{t\geq 2}(s_{t-1,q}r_{t-1,n})^{(k)}}. \quad (11)$$

$$\mu_q^{(k+1)} = \frac{\sum_t \sum_n (s_{t,q}r_{t,n})^{(k)} \mathbf{A}_n^{-1}\mathbf{x}_t}{\sum_t \sum_n (s_{t,q}r_{t,n})^{(k)}}. \quad (12)$$

$$\mathbf{\Sigma}_q^{(k+1)} = \frac{\sum_t \sum_n (s_{t,q}r_{t,n})^{(k)} (\mathbf{A}_n^{-1}\mathbf{x}_t - \mu^{(k+1)})(\mathbf{A}_n^{-1}\mathbf{x}_t - \mu^{(k+1)})^t}{\sum_t \sum_n (s_{t,q}r_{t,n})^{(k)}}. \quad (13)$$

### 3.3. MSE Transformation matrices estimation

The rotation matrices $\mathbf{A}_n$ allow to this family of models to a great degree of freedom as they can be any linear transformation for the feature vectors, therefore including the VTLN transformation naturally in the model and in this article we have focused on this transformations.

In order to estimate the transformation matrix as it has been shown in Section 3, we have followed the well known result of the multidimensional regression Minimum Square Error (MSE) criterion, which we sum up in this section. We have selected this data driven method as it is suitable for this task but also will provide the possibility of expanding to more transformations by changing the target data.

Let be a linear transformation for a $D$-dimensional source feature space samples $\mathbf{X}$ $(D \times T)$, where T is the number of samples, to a target feature space $\mathbf{Y}$ $(D \times T)$, and we define a general linear transformation as:

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}\mathbf{O}, \quad (14)$$

where we want to estimate $\mathbf{A}$ $(D \times D)$ and a bias term $\mathbf{b}$ $(D \times 1)$, being $\mathbf{O}$ an all ones $(1 \times T)$ matrix. Then we define the residual error $e$ to be the square sum of differences between the desired $\mathbf{Y}$ data and the projected data as,

$$e = Tr\left[(\mathbf{A}\mathbf{X} + \mathbf{b}\mathbf{O} - \mathbf{Y})(\mathbf{A}\mathbf{X} + \mathbf{b}\mathbf{O} - \mathbf{Y})^t\right]. \quad (15)$$

After taking derivatives with respect to $\mathbf{A}$ and $\mathbf{b}$ and equaling them to zero, we will obtain the well know result, valid if the mean in $t$ of $\mathbf{X}$ is $\mathbf{0}$:

$$\mathbf{A} = (\mathbf{X}\mathbf{X}^t)^{-1}\mathbf{X}\mathbf{Y}^t. \quad (16)$$

$$\mathbf{b} = (\mathbf{O}\mathbf{O}^t)^{-1}(\mathbf{Y} - \mathbf{A}\mathbf{X})\mathbf{O}^t. \quad (17)$$

In the experiments in this paper we have proposed a target data which is the VTLN warped feature vectors, $\mathbf{X}^\alpha$, in [5, 6], providing a transformation matrix for each one of the warping factors.

### 3.4. Search algorithm

After presenting the way that the model parameters are estimated, we now propose the search algorithm for decoding unlabeled sequences under this framework,

$$\phi_{q,n}(t) = \max_{n',q'} \{\phi_{q',n'}(t-1) \cdot \pi_{q',n',q,n}\} \cdot f(\mathbf{x}_t|q,n), \quad (18)$$

where $\phi()$ is the score state variable and $\pi$ vector contains the state transition probabilities and $f(\mathbf{x}_t|n,q)$ is the observation generation pdf described in (2).

This recursive expression is very similar to the one in [5] and the main difference is how the warping is done, since now is the model who tries to generate or evaluate the warped data instead of normalizing data to fit the model. In the new framework the covariance is normalized in the model description so the Jacobian normalization in [7] is included in the model. The same restrictions as in [5] have been applied to the transition matrix.

## 4. Results

In order to evaluate the performance of the new models, several experiments have been carried out. The task domain was isolated and connected digits in the TIDIGITS corpus, which is a noise free corpus organized in age and gender groups for a total of 326 speakers (111 men, 114 women, 50 boys, 51 girls). Since the main objective of the method is the speaker variability, this corpus and the proposed experimental method have been chosen.

In all the experiments 7 groups were defined in the training and testing partitions: 'boy', 'girl', 'man', 'woman', 'boy + girl', 'man + woman' and 'all'. For those 7 groups, HMM 16 state word models with increasing number of Gaussian components and a begin-end silence 3 states model and an inter word silence model of 1 state were trained. As feature set, the standard ETSI features plus the energy and their first and second derivatives, were used in all the experiments.

On the first experiment, the speaker variability reduction on a high mismatch task is tested, this experiment was performed on a subset of the corpus containing only isolated digits in order to test the ability of the proposed method to reduce inter speaker mismatch in a low training data availability context (3586 isolated digit utterances for training and 3582 for testing). Since mismatch conditions were tested, models for MATE where not reestimated and the simple expansion of the baseline pdfs was performed, a $20\%$ of deviation for $\alpha$ was set and $N = 5$. The results of the experiments are shown in Figure 1, where the WER is calculated as the mean of all the WERs of testing each of the defined group models with data from a different group (42 tests). It could be thought that MATE results comparison for a fixed number of Gaussian components as in [5] could carry an increase of the computation cost to obtain a performance that could also be achieved by increasing the number of Gaussian components, but we can observe in this experiment that the effect of overtraining is observable as the number of Gaussian components grow, since it is a small data set, and MATE can reduce the WER effectively under this kind of

Table 1: Recognition results in WER for each of the defined groups in the matched conditions.

| Group | Man | Woman | Boy | Girl | Man + Woman | Boy + Girl | All |
|---|---|---|---|---|---|---|---|
| baseline (% WER) | 1.79 | 1.04 | 1.90 | 1.17 | 2.00 | 1.50 | 2.67 |
| MATE (% WER) | 1.74 | 0.91 | 1.82 | 0.95 | 1.80 | 1.39 | 2.34 |
| % IMP | 2.8 | 12.7 | 4.2 | 18.9 | 9.8 | 7.4 | 12.3 |
| max deviation (%) | $\pm$10% | $\pm$10% | $\pm$5% | $\pm$10% | $\pm$15% | $\pm$5% | $\pm$15% |

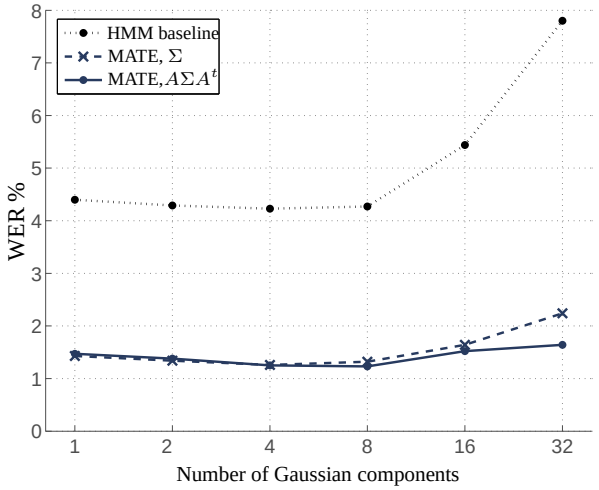


Figure 1: Mean WER in the speaker train and test models mismatch experiment for the baseline and MATE.

situations. Comparing the best result obtained in the HMM baseline, 4 Gaussians, the WER reduction then is a $70\%$ for MATE, (MATE std is $1.9\%$ vs. a $6.6\%$ for baseline), a $14\%$ of reduction was found for VTLN. This experiment shows the ability of generalization of MATE in speaker mismatch situations (e.g. test children with trained adult models). The results, $4\%$, for baseline are not surprising since HMMs are not able to generate/recognize samples of unseen data (group mismatch). In this experiment we have compared also the effect of the normalization of the covariance matrix as it has been described, as in a previous MATE [5] was not taken into account. Although the effect can be noticed, it appears that there not exists an statistically significative difference in our experiment. The second experiment includes the normalization as it does not require any additional computation cost.

In the second experiment, the group model and test examples are matched so that the variability inside the defined groups can be measured. It was evaluated for the complete TIDIGITS corpus size and results are presented in Table 1. As the corpus size is bigger in this experiment the effects of the overtraining have not been evaluated as in the previous one. The model parameters have been fixed to 1 Gaussian component per state in both baseline and MATE tests, $N = 5$ and various ranges of transformation factor have been evaluated in this case. MATE in this experiment was retrained with one iteration as it has been shown in Section 3. From the results it is interesting to note that for the best maximum range of transformation factor for the frequency axis, which is presented in the last row of the table, it is possible to check that, as expected, the more compact are the groups it tends to be smaller. For the more specialized groups: Man, Woman, Girl, Boy, the maximum transformation factor lays between 5% and 10%, and when they are merged them into bigger groups the best maximum range has increased up to a 15%. We have found WER reductions in all the groups (a 9.7% in mean), but the more significative and applicable to a real system are the last three columns, which correspond to the merged groups, as in many ASR systems there is no prior information if the kind of speaker.

## 5. Conclusions

In this paper we have presented a model for speech feature vector sequences in which it is believed to exist certain amount of local variability that the usual HMM framework is not able to model even the corpus size and the number of parameters is highly increased. The MATE includes naturally frame specific transformations of the speech in the state observation pdfs, by means of a linear projection and a kind of expansion of the states which does not increase substantially the number of parameters as they remain tied across the expansion.

The models have been tested on TIDIGITS corpus in two main experiments: speaker matched and speaker unmatched training and testing age-gender group conditions, obtaining good results in both of them, specially in the high mismatch cases in which the WER reduction can reach to a $70\%$ in a relatively small subset of the corpus and in the smaller mismatch experiments. When the model is trained with utterances coming from all the groups the WER reduction can be a $12.3\%$.

## 6. References


[1] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Trans. on SAP*, vol. 1, no. 6, pp. 49–60, 1998.

[2] D. Kim, S. Umesh, M. J. Gales, T. Hain, and P. Woodland, "Using VTLN for broadcast news transcription," in *Proc. ICSLP* , Jeju Island, S.Korea, October 2004.

[3] C. J. Legetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of the parameters of continuous density hidden markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[4] T. Fukada and Y. Sagisaka, "Speaker normalized acoustic modeling based on 3-D viterbi decoding," in *Proc. ICASSP* , Seattle, USA, 1998, vol. 1, pp. 437–440.

[5] A. Miguel, E. Lleida, R. Rose, L. Buera, and A. Ortega, "Augmented state space acoustic decoding for modeling local variability in speech," in *Proc. Eurospeech* , Lisbon, Portugal, 2005, pp. 3009–3012.

[6] R. Rose, A. Keyvani, and A. Miguel, "On the interaction between speaker normalization, environment compensation, and discriminant feature space transformations," in *Proc. ICASSP* , Toulouse, France, May 2006.

[7] M. Pitz and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," *IEEE Trans. on SAP*, vol. 13, no. 5, pp. 930–944, September 2005.

[8] A. Juan and E. Vidal, "Bernoulli mixture models for binary images," in *Proc. ICPR* , Cambridge, UK, August 2004, vol. 2.