Capturing Local Variability for Speaker Normalization in Speech Recognition

Antonio Miguel, Eduardo Lleida, *Member, IEEE*, Richard Rose, *Senior Member, IEEE*, Luis Buera, Óscar Saz, and Alfonso Ortega

***Abstract*—The new model reduces the impact of local spectral and temporal variability by estimating a finite set of spectral and temporal warping factors which are applied to speech at the frame level. Optimum warping factors are obtained while decoding in a locally constrained search. The model involves augmenting the states of a standard hidden Markov model (HMM), providing an additional degree of freedom. It is argued in this paper that this represents an efficient and effective method for compensating local variability in speech which may have potential application to a broader array of speech transformations. The technique is presented in the context of existing methods for frequency warping-based speaker normalization for ASR. The new model is evaluated in clean and noisy task domains using subsets of the Aurora 2, the Spanish Speech-Dat-Car, and the TIDIGITS corpora. In addition, some experiments are performed on a Spanish language corpus collected from a population of speakers with a range of speech disorders. It has been found that, under clean or not severely degraded conditions, the new model provides improvements over the standard HMM baseline. It is argued that the framework of local warping is an effective general approach to providing more flexible models of speaker variability.**

***Index Terms*—Automatic speech recognition (ASR), local warping, maximum likelihood, speaker normalization, vocal tract normalization.**

I. Introduction

SPEAKER variability has a negative impact on the performance of automatic speech recognition (ASR) systems. A speech modeling technique based on local spectral and temporal mismatch reduction is presented in this paper.

Local frequency variability is, to a limited extent, implicitly modeled by existing techniques such as hidden Markov model (HMM). In continuous observation density HMMs, there exists a basic mechanism to model the spectral variability that results from speaker-dependent variability in vocal tract shape. It is provided by the state dependent observation generating process, which is usually assumed to follow a multimodal probability density function (pdf) such as a Gaussian mixture model (GMM). The vocal tract shape deviations due to a large population of speakers can be captured by the state emission pdf as different components of the mixture. Consequently, a number of examples of vocal tract shapes are needed so that the components of the mixture can be estimated in the learning process. Therefore, multiple Gaussian components for each HMM state and speaker-matched training data are required in order to deal with this source of variability in a standard HMM, since the model itself cannot generalize speaker independent patterns.

Some methods have appeared in order to compensate the frequency axis variability including vocal tract length normalization (VTLN) [1], [2] which is a well-known method used for spectral warping. VTLN has been shown to be effective in compensating for long term average mismatch between the location of spectral peaks for a test speaker and the average spectral characteristics observed in the training samples. These average spectral characteristics can be difficult to characterize since training conditions are represented by the statistical HMM trained using utterances from a large population of speakers. More general methods exist for speaker adaptation such as maximum-likelihood linear regression (MLLR) [3], which reduce the mismatch between data and model. These regression procedures have limitations in their ability to retrain or adapt HMM models to a speaker. Usually, a large amount of speaker adaptation data and exact transcriptions are needed. In [4], a method for obtaining a frequency warping transformation using a constrained MLLR adaptation was shown, relating both methods. More recent approaches using a maximum likelihood linear transformation (MLLT) of the HMM to approximate the effects of spectral warping was proposed in [5]. The principal drawback of these methods is the need for previous speaker utterances or extra adaptation data in order to learn how to compensate for the spectral mismatch. In addition, all of these techniques attempt to estimate a single transformation which minimizes spectral mismatch over the entire utterance.

In the search for more flexible approaches there have been other approximations for modeling the frequency warping variability in the utterance. In [6], a model was presented (HMM2) in which the state-dependent speech observations were produced by a second Markov process, as the outputs of the hidden Markov process chain. This was intended to result in additional flexibility being provided for the frequency axis. In [7], an approach was presented for time–frequency modeling using a technique based on Markov random fields (MRFs). MRFs are an extension of HMMs in which the temporal index

Manuscript received December 21, 2006; revised October 25, 2007. This work was supported by the MEC of the Spanish goverment under national project TIN 2005-08660-C04-01. This work was supported in part by the National Sciences and Engineering Research Council of Canada program number 307188-2004. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Abeer Alwan.

A. Miguel, E. Lleida, L. Buera, O.Saz, and A. Ortega are with the Department of Electronic Engineering and Communications, University of Zaragoza, 50009 Zaragoza, Spain (e-mail: amiguel@unizar.es; lleida@unizar.es; lbuera@unizar.es; oskarsaz@unizar.es; ortega@unizar.es).

R. Rose is with the Department of Electrical and Computer Engineering, McGill University, Montreal, H3A 2T5 QC, Canada (e-mail: rose@ece.mcgill.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2007.914114

assigned to states in decoding is substituted by a spatial one, providing dependency relationships on the form of a 2-D neighborhood. In that work, local inter-band and inter-frame interactions present in speech were investigated.

The main motivation in this work is to find a normalization model able to capture speaker variability. The method described in this work provides a mechanism for the VTLN spectral warping procedure to be locally optimized. The model tries to capture local frequency deformations of the spectrum envelope, which are known to originate in physiological differences among vocal tract dimensions and variation in articulatory trajectories. A more complex and flexible speech production scheme can be assumed, in which local elastic deformations of the speech can be captured or generated by the model. The method proposed, referred to here as the augMented stAte space acousTic dEcoder (MATE), includes both training and search algorithms, which attempt to overcome the limitations of existing methods for compensating for spectral variability in ASR. MATE consists of an expansion of the HMM state space to provide local transformations that are estimated as part of a dynamic programming search through this augmented state space.

The first approaches to this paradigm were proposed in [8] and then followed by [9], [10] in a more general framework allowing the procedure to additionally be applied to local time axis optimization. In [11], the interaction between this procedure and methods to increase class discrimination and noise robustness were investigated. More recently, in [12], the effect of Jacobian normalization through the estimation of a linear transformation determinant was investigated. In [13], the performace of the MATE technique was evaluated on a Spanish language corpus collected from a population of speakers with a range of speech disorders [14], providing promising results.

There have been many previous approaches to augmenting the HMM state space to model sources of variability in ASR [15], [16] or in confidence measures [17]. These include attempts to model extrinsic sources of variability. For example, an expanded HMM state space and a modified Viterbi algorithm were defined to model the joint probability of speech and noise in HMM model decomposition [15], [16]. The approaches presented in this paper can be better described as attempts to model intrinsic sources of variability. This is accomplished by expanding the HMM state space to characterize spectral and temporal variability in speech production by defining distinct states to represent degrees of spectral and temporal warping. Hence, these approaches may be considered to be more closely related to techniques that use graphical models to augment the state space to define states that represent underlying articulatory events [18].

There are several issues addressed in this paper relating to the MATE formalism. MATE will be presented as an expansion of the HMM state space in which the optimum frame-specific warping functions are chosen to maximize the global path likelihood in the Viterbi search. This procedure requires only a single pass over the input utterance and produces frame-specific estimates of the warping functions. Training and decoding algorithms for MATE framework are presented in this paper.

This paper is organized as follows. Section II reviews VTLN, a frequency warping approach to speaker normalization, which is one of the existing techniques used for speaker mismatch reduction. Section III presents the model formulation, the modified Viterbi search algorithm performed in the augmented state space, and the procedure for estimating the parameters using the EM algorithm. Section IV describes the procedure for estimation of dynamic cepstra using frame specific temporal resolution and how this procedure is implemented with augmented state space models. Section V presents the results of the experimental study describing the performance of these new algorithms with respect to the performance of previous more well known existing techniques. Finally, discussion and conclusions are presented in Section VI.

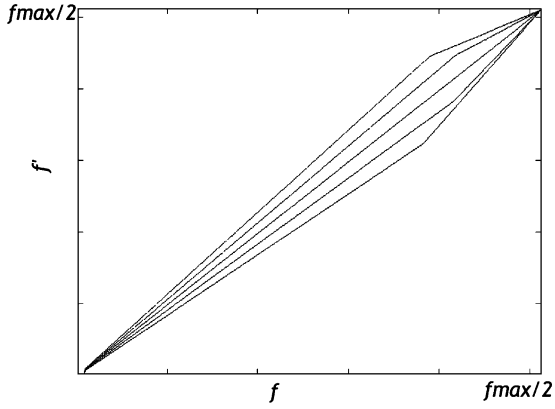

Fig. 1. Set of piecewise linear frequency warping functions $g^{\alpha}(f)$, where fmax is the sampling frequency.

II. Maximum-Likelihood Frequency Warping

In frequency warping-based speaker normalization techniques, such as VTLN, a warped frequency scale is produced

$$f' = g^{\alpha}(f) \tag{1}$$

where f is the original frequency and f' is the warped frequency, by selecting an optimum warping function from an ensemble of N linear frequency warping functions, $\mathcal{G} = \{g^{\alpha_n}(\cdot)\}_{n=1}^{N}$, as described in [2]. The warping functions are usually of the form illustrated by the curves in Fig. 1. The optimum warping factor $\alpha_{n'}$ is chosen to maximize the average likelihood of a T length sequence of frequency-warped cepstrum observation vectors $\mathbf{X}^{\alpha_n} = \{\mathbf{x}_t^{\alpha_n}\}_{t=1}^{T}$ with respect to an HMM and $n = 1, \ldots, N$.

A. Front-End Feature Extraction Notation

Let us briefly describe the procedure for extracting the acoustic features following, for convenience in a matrix form notation. The set of M component power spectral magnitude values for a frame t of an utterance are represented as an M dimensional vector $\mathbf{p}_t$. The B channel filter bank is applied to the M spectral magnitude values. It is implemented as a set of B triangular weighting functions. It is represented in matrix form as a $B \times M$-dimensional matrix $\mathbf{F}$.

The logarithm of the outputs of these filters $\mathbf{o}_t$ are computed as $\mathbf{o}_t = \log(\mathbf{F} \cdot \mathbf{p}_t)$. Then, the C component cepstrum vectors $\mathbf{c}_t$ are computed as the discrete cosine transform of $\mathbf{o}_t$, $\mathbf{c}_t = \mathbf{W}_F \cdot \mathbf{o}_t$. The $C \times B$ matrix $\mathbf{W}_F$ corresponds to the initial C vectors of the discrete cosine transform basis of dimension

B, excluding the zeroth order term, which provides a frequency projection.

In order to obtain more discriminative and robust feature vectors, dynamic features are usually included. This is assumed to partially solve the observation independence assumption of HMMs [19]. A temporal sliding window of length L is taken over the cepstrum sequence, being L the length of the cepstrum window over which we compute the dynamic features. We can build an $L \times L'$ time projection matrix $\mathbf{W}_T$ which provides a temporal projection of the features window, which traditionally for the speech community takes the form of the static cepstrum coefficients and their first and second derivatives, so that $L' = 3$.

To compute the dynamic features, we construct the $\mathbf{O}_t$ matrix, appending L filter bank output frames centered on frame t

$$\mathbf{O}_t = \left\{\mathbf{o}_{t-\lfloor L/2 \rfloor}, \ldots, \mathbf{o}_t, \ldots, \mathbf{o}_{t+\lfloor L/2 \rfloor}\right\}. \tag{2}$$

Then, the dynamic features are calculated as

$$\mathbf{x}_t = \mathrm{vec}(\mathbf{W}_F \cdot \mathbf{O}_t \cdot \mathbf{W}_T) \tag{3}$$

where $\mathbf{x}_t$ is the feature vector, and $\mathrm{vec}(.)$ is a matrix to vector operator, which converts the $C \times 3$ matrix to a column vector. The notation used to describe feature extraction given in (3) is important because this formulation allows us to present both temporal and frequency local projections. The class of local transformations considered in this work consists of either modifying the frequency projection or the temporal projection for a single frame. The final step is to append to this vector the energy parameter and their derivatives which can be computed by applying the same temporal projection $\mathbf{W}_T$ matrix.

B. Frequency-Warped Features

The procedure for obtaining the frequency-warped cepstrum sequence $\mathbf{X}^\alpha$ involves modifying the Mel-frequency filter bank $\mathbf{F}$ that is used in Mel-frequency cepstrum coefficient (MFCC) feature analysis [2]. Following previous notation, frequency warping can be implemented by applying the warping functions $g^\alpha(\cdot)$ to the array of filter bank coefficients in the matrix $\mathbf{F}^\alpha$. The set of warped cepstrum vectors for a frame t can then be expressed as

$$\mathbf{c}_t^\alpha = \mathbf{W}_F \cdot \mathbf{o}_t^\alpha = \mathbf{W}_F \cdot \log(\mathbf{F}^\alpha \cdot \mathbf{p}_t). \tag{4}$$

An alternative implementation can be used to obtain $\mathbf{c}_t^\alpha$. In [4], it has been shown that, with the proper assumptions, the filter bank modification in $\mathbf{F}^\alpha$ is equivalent to applying a linear transformation in the cepstrum domain

$$\mathbf{c}_t^\alpha = \mathbf{A}^\alpha \cdot \mathbf{c}_t. \tag{5}$$

Then, the discrete Fourier transform (DCT) frequency projection $\mathbf{W}_F$ and the linear projection $\mathbf{A}^\alpha$ can be unified in a local frequency warping matrix

$$\mathbf{x}_t^\alpha = \mathrm{vec}\left(\mathbf{W}_F^\alpha \cdot \mathbf{O}_t \cdot \mathbf{W}_T\right) \tag{6}$$

where $\mathbf{W}_F^\alpha = \mathbf{A}^\alpha \cdot \mathbf{W}_F$.

C. Maximum-Likelihood Estimation

In standard frequency warping methods, the optimal transformation factor $\alpha_{n'}$ within the discrete set $\mathcal{A} = \{\alpha_n\}_{n=1}^N$ for a speaker is obtained from a maximum-likelihood estimation process involving speech samples and transcriptions [2]. Following the warping process described previously, warped features $\mathbf{X}^{\alpha_n}$ are computed for each warping factor n. Then, the optimal warping factor $\alpha_{n'}$ is obtained by maximizing the likelihood of the warped utterance with respect to the model parameters $\boldsymbol{\Theta}^{(0)}$ and the transcription $\mathbf{W}$

$$\alpha_{n'} = \arg\max_{\alpha_n} \left\{ p\left(\mathbf{X}^{\alpha_n} | \boldsymbol{\Theta}^{(0)}, \mathbf{W}\right)\right\}. \tag{7}$$

While this maximum-likelihood-based warping procedure has been shown to significantly reduce word error rate (WER) in many cases, it has two important limitations. The first one is that it can be unwieldy to apply. 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, when the transcription $\mathbf{W}$ is unknown. This initial word string is then used in a second pass to compute the likelihood of the N-warped utterances by aligning $\mathbf{X}^{\alpha_n}$ with the decoded word string.

The second limitation is related to the fact that only a single linear warping function is selected for an entire 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, this technique estimates a single transformation for an entire utterance. The procedure described in this work addresses both of these issues.

III. Augmented State Space Models

Augmented state space models are presented here for modeling variations in vocal tract shape that occurs during speech utterances and across speakers. A new degree of freedom is added to track those local changes in a HMM framework. Since this paper presents the model from a feature normalization point of view, the formulation, EM derivation, and decoding algorithms expand the work presented in [10]. This section presents the description of the formulation for training normalized models under the augmented state space framework. Also, a modified search algorithm for decoding speech utterances is presented since in MATE, the Viterbi algorithm is implemented in an augmented state space which allows frame-specific spectral warping functions to be estimated as part of the search.

Fig. 2 shows the local alignment basic idea, by means of three examples of utterances (manually synthesized and modified for illustrative purposes) of the same phoneme transition (plosive-vowel) for different speakers. The three spectrograms along the top row of Fig. 2 illustrate how the distribution of spectral energy for the plosive is consistent across all three speakers. However, there is considerable variability in the formant positions for the vowel across speakers. The figure illustrates the potential benefits of selectively warping the vowel portion of these utterances to be more consistent with the average spectrum of the HMM model shown on the right. In this example, estimating a single warping transformation over the whole utterance would not have the same effect. The local alignment

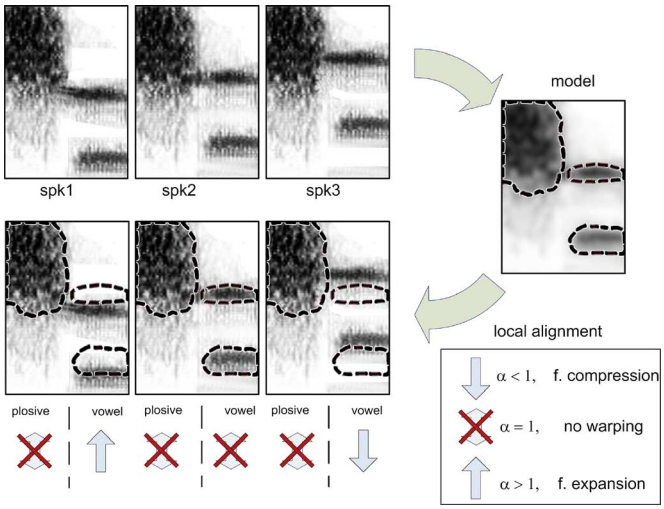

Fig. 2. Example of local alignment, observations are transformed towards the model, where speech data has been manually synthesized and modified for illustrative purposes.

procedure works in the direction indicated by the arrows, compressing or expanding the frequency scale. It is applied only in the sounds where there is significant variation with respect to the HMM model, in this case in the vowel. The main motivation for MATE is to locally adapt speech features to the general model.

Since, it is necessary for MATE to track local deformations of speech, the simpler solution is to define a discrete set of candidate warpings in a manner similar to that used in VTLN [11]. Then a second criterion adopted is to impose some inertia and speed constraints to the local transformation dynamics. Those constraints allow us to assume a model for the warping dynamics which follows a first-order Markov process, restricting the maximum physical speed they can change and providing a dependency on the previous warping. The MATE solution is an expansion of the state space, so that expanded states are identified with warpings, and we can apply the same dynamic constraints associated with the standard state sequence in HMM modeling. Therefore, the expanded models fit naturally in the HMM framework.

Let us describe in detail the augmented state space model for a discrete set of N local transformations. Given a Q state HMM, the MATE state expansion is defined so that N expanded states are generated from each of the Q original model states as the Cartesian product, where the state identifiers of original model (q) now become pairs (q,n). The expanded state space size is $Q \times N$. An interesting consequence is that the sequences of deformations can be learnt in the model parameters as transition probabilities between the expanded states, i.e., transition from a state (q,n) to (q',n').

To indicate the state producing an observation, let us define the indicator vector variables $\mathbf{s}_t$ and $\mathbf{r}_t$ for the state q and the transformation n for a time index t, respectively. The vector $\mathbf{s}_t \in \{0,1\}^Q$ indicates the state index q which generated the observation at time index t with $s_{t,q}=1$ and zeros elsewhere, as in [12] and [20]. The vector $\mathbf{r}_t$ is another indicator vector $\mathbf{r}_t \in \{0,1\}^N$, with $r_{t,n}=1$ for a local transformation index n at time t and zeros elsewhere.

As in [4] and [21], we tie the pdf of the augmented state (q,n) to the unwarped state (q) pdf as

$$p(\mathbf{x}_t|s_{t,q}=1,r_{t,n}=1)=p\left(f_{\mathbf{r}_t}(\mathbf{x}_t)|s_{t,q}=1\right)\cdot\left|\frac{\delta f_{\mathbf{r}_t}(\mathbf{x}_t)}{\delta \mathbf{x}_t}\right| \quad (8)$$

$$=p\left(\mathbf{x}_t^{\alpha_n}|s_{t,q}=1\right)\cdot J(n) \quad (9)$$

where $p(\cdot|s_{t,q}=1,r_{t,n}=1)$ is the augmented state observation pdf, and $p(\cdot|s_{t,q}=1)$ is the original model state observation pdf. The warped feature vector $\mathbf{x}_t^{\alpha_n}$ is obtained from (6), and n is the warping factor at time index t. Therefore, the state space expansion operation does not increase the state observation pdf parameters which are tied to their original state q. The additional factor $J(n)$ is the determinant of the Jacobian of the transformation n

$$J(n)=\left|\frac{\delta f_{\mathbf{r}_t}(\mathbf{x}_t)}{\delta \mathbf{x}_t}\right| \quad (10)$$

which is discussed in Appendix I for the speech dynamic observation vectors $\mathbf{x}_t$. To summarize, the augmented state space model can be seen as a 2-D HMM topology of size $Q \times N$ with a pdf for each state (q,n) tied to the original 1-D HMM state (q) pdf.

The augmented model can be interpreted within the dynamic Bayesian networks framework, as it can be done with standard HMM [18], [22]. The model topology would be an expansion of a single HMM Bayesian network topology to represent augmented states (q,n), which would generate warped feature vectors $\mathbf{x}_t^{\alpha_n}$. Additional arcs between augmented states would be required to model MATE state transition probabilities. There is also another point of view for the interpretation of MATE which can also be thought to be a kind of composed model. The model composition could be defined similarly to [15] and [16], with two separate models, one for the standard word HMM and another for the warping factors.

A. Complete Model

In order to define the local transformation estimation process, let us first 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. Then, let $\mathbf{X}=\{\mathbf{x}_1,\ldots,\mathbf{x}_T\}=\{\mathbf{x}_t\}_1^T$ be a T length sequence of speech observations, a state label sequence, $\mathbf{S}=\{\mathbf{s}_1,\ldots,\mathbf{s}_T\}=\{\mathbf{s}_t\}_1^T$, a frame specific transformation label sequence, $\mathbf{R}=\{\mathbf{r}_1,\ldots,\mathbf{r}_T\}=\{\mathbf{r}_t\}_1^T$. We can produce a frame-specific warped feature vector sequence as $\hat{\mathbf{X}}=\{f_{\mathbf{r}_1}(\mathbf{x}_1),\ldots,f_{\mathbf{r}_T}(\mathbf{x}_T)\}$, as described in (6), where $\mathbf{x}_t \in \mathbb{R}^D$ (with D the dimension of the feature vector) and $\mathbf{s}_t$ and $\mathbf{r}_t$ are the augmented state labels for time index t as previously defined.

The joint pdf of a local warped sequence of this kind can be written as follows using Bayes rule

$$(\mathbf{X},\mathbf{S},\mathbf{R})=p(\mathbf{S},\mathbf{R})\cdot p(\mathbf{X}|\mathbf{S},\mathbf{R}) \quad (11)$$

$$=\prod_{t=1}^{T} p\left(\mathbf{s}_t,\mathbf{r}_t|\{\mathbf{s}_t\}_1^{t-1},\{\mathbf{r}_t\}_1^{t-1}\right)\cdot\prod_{t=1}^{T} p\left(\mathbf{x}_t|\{\mathbf{x}_t\}_1^{t-1},\mathbf{S},\mathbf{R}\right) \quad (12)$$

which is computationally intractable.

Assuming first-order Markov assumptions over (12) and assuming independence between observations, we can approximate it by a simpler expression

$$p(\mathbf{X},\mathbf{S},\mathbf{R}) \simeq \prod_{t\geq 1} p(\mathbf{s}_t,\mathbf{r}_t|\mathbf{s}_{t-1},\mathbf{r}_{t-1}) \cdot \prod_{t\geq 1} p(\mathbf{x}_t|\mathbf{s}_t,\mathbf{r}_t) \quad (13)$$

where the inertia and memory constraints imposed on the dynamics of the local transformations can be identified in the term $p(\mathbf{s}_t,\mathbf{r}_t|\mathbf{s}_{t-1},\mathbf{r}_{t-1})$, which will be an important part of the search algorithm. Then, as in a standard HMM, the state indicator vectors follow a Multinomial distribution of parameters

$$\mathbf{\Pi} = \{\pi_{q,n,q',n'}\}_{q=1,n=1,q'=1,n'=1}^{Q,N,Q,N} \quad (14)$$

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

$$\pi_{q,q',n,n'} = p(s_{t,q'}=1, r_{t,n'}=1|s_{t-1,q}=1, r_{t,n}=1). \quad (15)$$

The joint probability in (13) can be written as in (16).

Taking into account that the indicator variables are zeros in all positions except one, then we can express (13) as (16), where the expanded state (q,n) pdf for the warped data $p(\cdot|s_{t,q}=1, r_{t,n}=1)$ follows a distribution of the form of (8). The ensemble of parameters composed by $\mathbf{\Pi}$ and the state pdfs are the parameter set, referred as $\mathbf{\Theta}$

$$\begin{aligned} p(\mathbf{X},\mathbf{S},\mathbf{R}) = & \prod_{q,n} [\pi_{0,0,q,n}]^{s_{1,q}r_{1,n}} \\ & \cdot \prod_{t\geq 2} \prod_{q,q',n,n'} [\pi_{q,n,q',n'}]^{s_{t-1,q}r_{t-1,n}s_{t,q'}r_{t,n'}} \\ & \cdot \prod_{t\geq 1} \prod_{q,n} [p(\mathbf{x}_t^{\alpha_n}|s_{t,q}=1)\cdot J(n)]^{s_{t,q}r_{t,n}}. \end{aligned} \quad (16)$$

B. MATE Decoding Algorithm

A Viterbi beam search decoder for continuous speech recognition is implemented by propagating paths into the nodes of a 2-D trellis. Each node of the trellis corresponds to one of the Q HMM states $\{(q)\}_1^Q$ evaluated for observation vectors $\mathbf{x}_t, t=1,\ldots,L$. In the MATE decoder, the state space is expanded by a factor of N, where N is the size of the ensemble of warping functions. This effectively results in a 3-D trellis. Each node of this augmented trellis now corresponds to the augmented state space defined in Section III.

The optimum sequence of states is identified for the decoding process in a standard HMM using the Viterbi algorithm, which we express as

$$\phi_q(t) = \max_{q'} \{\phi_{q'}(t-1)\cdot \pi_{q',q}\} \cdot p(\mathbf{x}_t|s_{t,q}=1) \quad (17)$$

where $\phi_q(t)$ is the likelihood of the optimum path terminating in HMM state q at time t, and $\pi_{q',q}$ is the transition probability from state (q') to state (q). The $\max$ is computed over all states that are permitted by the HMM model to propagate into state (q) which, for a left-to-right HMM topology, would be $(q-1)$ and (q). In addition to the accumulated path likelihood $\phi_q(t)$, a state traceback variable $\varphi_q(t)$ is defined which stores the sequence of state indices identified in (17).

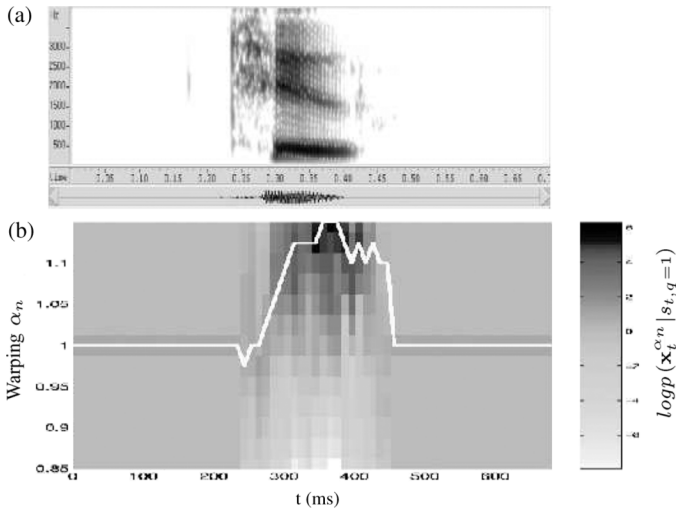

Fig. 3. (a) Spectrogram of an example of the word "two." (b) Likelihood of observation of each transformation factor for the states of the best path found by the Viterbi algorithm (global VTLN warping $\alpha = 1.13$) and optimal path in white line.

In the MATE decoder, the optimum sequence of states in the augmented state space is identified using a modified Viterbi algorithm, where the accumulated path likelihood in the augmented state space $\phi_{q,n}(t)$ is defined recursively as

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

where $\phi_{q,n}(t)$ is the likelihood of the optimum path terminating in state (q,n) at time t, and $\pi_{q',n',q,n}$ is the transition probability from state (q',n') to state (q,n). The $\max$ is computed over all states that are permitted by the HMM model to propagate into state (q,n), and the observation pdfs in the augmented state space share the same parameters as the original states as expressed in (8). Also, an augmented state space traceback variable $\varphi_{q,n}(t)$ is defined to store the sequence of state indices identified in (18).

Structural constraints can be placed on standard HMM topologies by constraining the allowable transitions between HMM states. Constraints can also be placed on the transformations g^{α_n} that are permitted at state (q,n) in the augmented state space. These constraints can be applied by setting a subset of the transition probabilities $\pi_{q,n,q',n'}$ equal to zero. In this paper, transition probabilities were constrained so that the frequency-warping transformations applied to adjacent frames were required to be taken from adjacent indices in the ensemble $\mathcal{G}$

$$\pi_{q',n',q,n} = 0, \text{ if } |n'-n| > 1. \quad (19)$$

These constraints have the effect of reducing the computational complexity in search. Furthermore, they also provide a mechanism for limiting the degrees of freedom in the application of spectral transformations to reflect a more physiologically plausible degree of variability and reduce the probability of local alignment errors between different spectrally close sounds. Additional constraints can be applied. For example, HMM states for nonspeech models are constrained so that no transformations

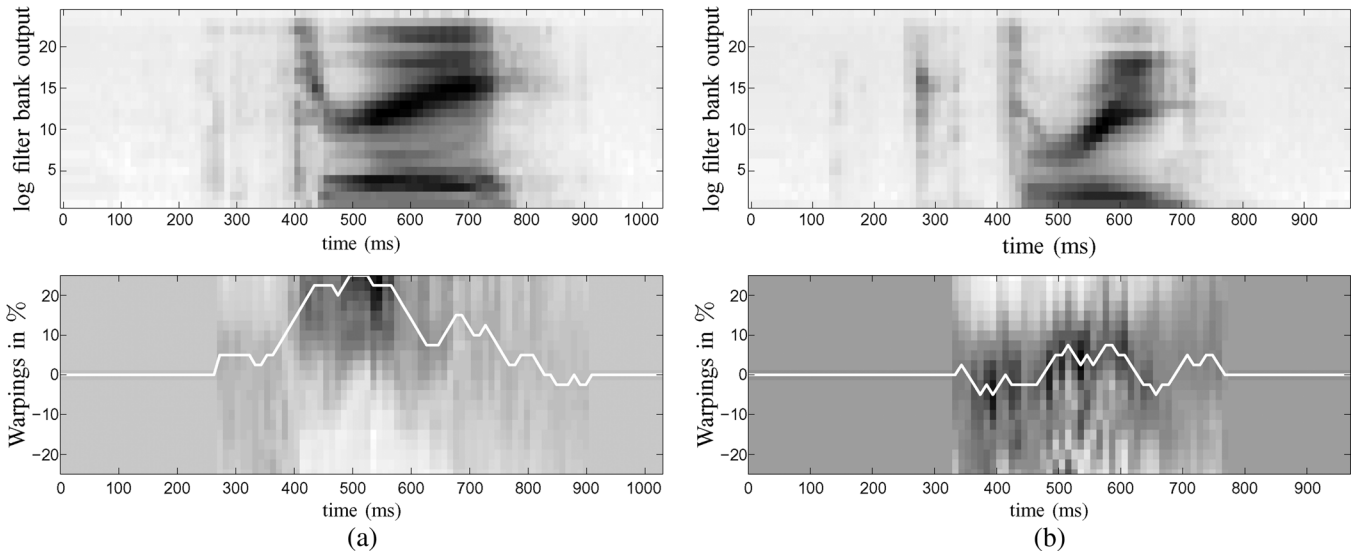

Fig. 4. Examples of the best warping decoded sequence for two utterances of the digit "three" with respect to an HMM trained from a population of adult speakers. (a) Uttered by a boy. (b) Uttered by a man.

can be applied to the observation vectors that are associated with those states.

An illustration of the effect of the MATE decoder is provided by the spectrogram and log likelihood plots for an utterance of the word "two" shown in Fig. 3(a) and (b), respectively. Fig. 3(b) displays the log likelihoods plotted versus time for a set of $N = 21$ possible frequency warping transformations that correspond to compressing or expanding the frequency axis by as much as 15%. This plot was obtained by performing Viterbi decoding on this utterance using the modified Viterbi algorithm given in (18). The trajectory of warping functions corresponding to the optimal path is indicated by a superimposed white line. It is interesting that, for the initial part of the word corresponding to the unvoiced aspirated "t" sound, values of $\alpha \simeq 1.00$ are chosen. This effectively corresponds to the choice of no spectral transformation for this region. However, in the voiced part of the word, a warping value is chosen that is similar to the value selected by the global linear VTLN warping factor that had been estimated for this utterance $(\alpha = 1.13)$.

Figs. 4 and 5 compare the local warping factors selected for utterances of a given word that are taken from speakers that are either closely matched or highly mismatched to the training speaker population. The plot on the top of each pair of plots in Figs. 4 and 5 depicts a spectrogram of MFCC log filter bank energies. The plot on the bottom corresponds to the same trajectory of warping factors shown in Fig. 3(b). The figures show the best found sequence of warping factors for a case of matched conditions (adult utterances decoded with an adult model), and mismatched conditions (children utterances decoded with the adult model). We can notice there how due to the physical restriction in the speed of change imposed by the constraint (20), the algorithm finds a smooth best path among all the likelihood values. Also, we can notice that in the case of matched conditions, there is no need for warping and in the case of the children utterances more warping effort has to be made. We also can see again that the most warped frames correspond to voiced sounds in all the utterances. In both figures it is clear that, in the case of children, formant centers are located at higher frequencies than in adults, since the center frequency of the resonances of the vocal tract are inversely proportional to the length of the vocal tract. Since the HMM models where trained from an adult speaker population, it is not surprising that the margnitude of the warping factors selected for the children speakers in Figs. 4(a) and 5(a) is greater than the magnitude of the warping factors chosen for the adults. These anecdotal examples will be supported by more vigorous ASR results in Section V.

The work in [8] also addressed the extension of the search space in ASR by estimating local frequency-warping parameters. However, their approach is limited to the definition of a decoding framework for local warping factor estimation. This section expands on the system developed in [8] by developing a general modeling framework that enable the following capabilities. First, all parameters in the new model can be estimated from data. Additional domain knowledge can be applied by constraining the structure of this general model rather than making ad hoc assumptions from the outset. For example, it was assumed in [8] that transition probabilities were constrained in the following way:

$$\pi_{q',n',q,n} \simeq \pi_{q',q} \cdot \pi_{n',n} \tag{20}$$

with $\pi_{n',n}$ defined as an indicator function with values 0 and 1. However, in MATE, the expanded state transitions, i.e., (q', n') to (q, n) can be estimated in training. Second, an exact solution to likelihood computation from transformed observations can be implemented in this framework by incorporating the Jacobian term in local probability estimation. Finally, the general framework facilitates the use of the augmented state space to represent other sources of variability including temporal compression and expansion as described in Section IV.

C. EM Training Algorithm

Since the label sequences $\mathbf{S}$ and $\mathbf{R}$ are not observable, it is not possible to solve for the parameters of the augmented state

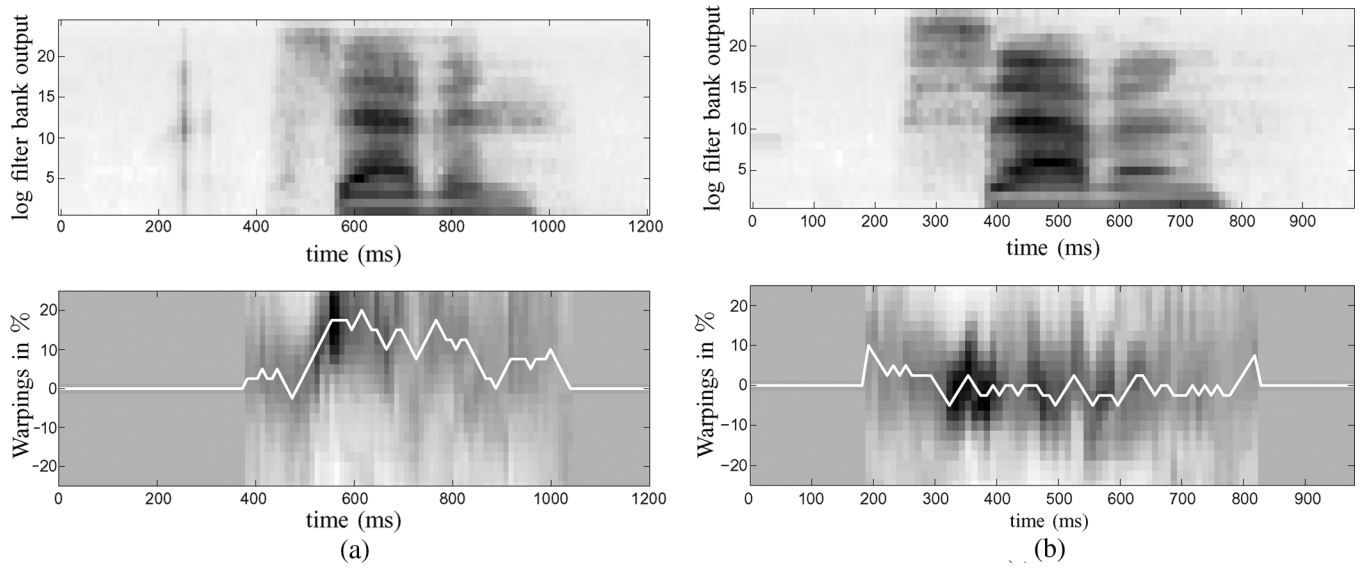

Fig. 5. Examples of the best warping decoded sequence for two utterances of the digit "seven" with respect to an HMM trained from a population of adult speakers. (a) Uttered by a girl. (b) Uttered by a woman.

space model directly. Treating $\mathbf{S}$ and $\mathbf{R}$ as hidden variables, the parameters of the augmented state space model can be estimated by using the expectation-maximization (EM) algorithm [23]. The development of the EM algorithm for MATE is similar to that used for standard continuous-density HMM [24].

The first step, the E expectation step, consists of calculating the auxiliary function $Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(k)}) = E[\log p(\mathbf{X},\mathbf{S},\mathbf{R}|\boldsymbol{\Theta})|\mathbf{X},\boldsymbol{\Theta}^{(k)}]$, which 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 (21) for our model, where the expressions noted as $(\cdot)^{(k)}$ refer to the expected values of the variable between the parentheses

$$\begin{aligned} Q\left(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(k)}\right) &= \sum_{q,n}(s_{1,q}r_{1,n})^{(k)}\cdot\log\pi_{0,0,q,n} \\ &+\sum_{t\geq 2}\sum_{q,q',n,n'}(s_{t-1,q}r_{t-1,n}s_{t,q'}r_{t,n'})^{(k)} \\ &\cdot\log\pi_{q,n,q',n'} \\ &+\sum_{t\geq 1}\sum_{q,n}(s_{t,q}r_{t,n})^{(k)}\cdot\log\left[p(\mathbf{x}_t^{\alpha_n}|s_{t,q}=1)\right. \\ &\cdot J(n). \end{aligned} \tag{21}$$

In the first one, the expected value for a state and a transformation at a time t can be calculated as

$$\begin{aligned} (s_{t,q}r_{t,n})^{(k)} &= E\left[s_{t,q}r_{t,n}|\mathbf{X},\boldsymbol{\Theta}^{(k)}\right] && (22)\\ &= \sum_{\forall t,q,n} s_{t,q}r_{t,n} \\ &\quad\cdot p\left(s_{t,q}=1,r_{t,n}=1|\mathbf{X},\boldsymbol{\Theta}^{(k)}\right) && (23)\\ &= p\left(s_{t,q}=1,r_{t,n}=1|\mathbf{X},\boldsymbol{\Theta}^{(k)}\right) && (24)\\ &= \frac{p\left(\mathbf{X},s_{t,q}=1,r_{t,n}=1|\boldsymbol{\Theta}^{(k)}\right)}{p\left(\mathbf{X}|\boldsymbol{\Theta}^{(k)}\right)} && (25) \end{aligned}$$

since all terms of the sum in (23) are 0 except the one pointed by the indicator vector components $s_{t,q}$ and $r_{t,n}$.

In the second one, the expected value of a transition

$$\begin{aligned} &(s_{t-1,q}r_{t-1,n}s_{t,q'}r_{t,n'})^{(k)} && (26)\\ &= E\left[s_{t-1,q}r_{t-1,n}s_{t,q'}r_{t,n'}|\mathbf{X},\boldsymbol{\Theta}^{(k)}\right] && (27)\\ &= \sum_{\forall t,q,n,q',n'} s_{t,q}r_{t,n}s_{t,q'}r_{t,n'} && (28)\\ &\quad\cdot p\left(s_{t-1,q}=1,r_{t-1,n}=1,s_{t,q'}=1,r_{t,n'}=1|\mathbf{X},\boldsymbol{\Theta}^{(k)}\right) && (29)\\ &= p\left(s_{t-1,q}=1,r_{t-1,n}=1,s_{t,q'}=1,r_{t,n'}=1|\mathbf{X},\boldsymbol{\Theta}^{(k)}\right) && (30)\\ &= \frac{p\left(\mathbf{X},s_{t-1,q}=1,r_{t-1,n}=1,s_{t,q'}=1,r_{t,n'}=1|\boldsymbol{\Theta}^{(k)}\right)}{p\left(\mathbf{X}|\boldsymbol{\Theta}^{(k)}\right)} && (31) \end{aligned}$$

where all the terms of the sum are also 0 except the indicated one, corresponding to states (q,n) and (q',n').

Those expressions are difficult to calculate directly, but thanks to the expanded auxiliary forward and backward functions $a_{q,n}(t)$, $b_{q,n}(t)$, which can be calculated recursively, computations are reduced to an affordable level as in standard HMM. The definition of those variables and the calculation of the expected values using them is formulated in Appendix II.

The second step M, the maximization step, consists of maximizing the $Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(k)})$ function with respect to the model parameters at iteration k to obtain the values for the parameters in the next iteration subject to the constraint

$$\sum_{q',n'}\pi_{q,n,q',n'} = 1 \tag{32}$$

for all $q=1,\ldots,Q$ and $n=1,\ldots,N$.

After computing the auxiliary function and the expected values, the new parameters will be obtained from the optimization

$$\boldsymbol{\Theta}^{(k+1)} = \arg\max_{\boldsymbol{\Theta},\boldsymbol{\lambda}} Q\left(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(k)}\right) - \sum_{q,n}\lambda_{q,n}\left(\sum_{q',n'}\pi_{q,n,q',n'} - 1\right) \quad (33)$$

for all $q = 1,\ldots,Q$ and $n = 1,\ldots,N$. Taking derivatives with respect to the parameter set $\boldsymbol{\Theta}$ and the Lagrange multipliers and equating them to zero, we obtain the following expressions for the parameter estimations in the iteration $k+1$ $\boldsymbol{\Theta}^{(k+1)}$ as:

$$\pi_{q,n,q',n'}^{(k+1)} = \frac{\sum_{t\geq 2}(s_{t-1,q}r_{t-1,n}s_{t,q'}r_{t,n'})^{(k)}}{\sum_{t\geq 2}(s_{t-1,q}r_{t-1,n})^{(k)}} \quad (34)$$

for all possible values of (q,n) and (q',n') over the augmented state space.

If the state observation pdf is a mixture of Gaussians

$$p(\mathbf{x}_t|s_{t,q}=1) = \sum_{c=1}^{C} w_c \cdot g(\mathbf{x}_t|s_{t,q}=1, z_{t,c}=1) \quad (35)$$

where C is the number of components, $z_{t,c}$ is a hidden variable indicating the c component generating the observation and $g(\mathbf{x}|s_{t,q}=1,z_{t,c}=1) \sim \mathcal{N}(\boldsymbol{\mu}_{c,q},\boldsymbol{\Sigma}_{c,q})$. The expected value for the hidden variable is

$$(z_{t,c})^{(k)} = \frac{w_c^{(k)} g\left(\mathbf{x}_t^{\alpha_n}|s_{t,q}=1,z_{t,c}=1\right)}{\sum_{c'} w_{c'}^{(k)} g\left(\mathbf{x}_t^{\alpha_n}|s_{t,q}=1,z_{t,c'}=1\right)}. \quad (36)$$

Then, the GMM parameter estimation expressions are

$$w_c^{(k+1)} = \frac{\sum_t\sum_{q,n}(s_{t,q}r_{t,n})^{(k)}(z_{c,t})^{(k)}}{\sum_t\sum_{q,n}\sum_{c'}(s_{t,q}r_{t,n})^{(k)}(z_{c',t})^{(k)}} \quad (37)$$

$$\boldsymbol{\mu}_{q,c}^{(k+1)} = \frac{\sum_t\sum_n(s_{t,q}r_{t,n})^{(k)}(z_{c,t})^{(k)}\mathbf{x}_t^{\alpha_n}}{\sum_t\sum_n(s_{t,q}r_{t,n})^{(k)}(z_{c,t})^{(k)}} \quad (38)$$

$$\boldsymbol{\Sigma}_{q,c}^{(k+1)} = \quad (39)$$

$$\frac{\sum_t\sum_n(s_{t,q}r_{t,n})^{(k)}(z_{c,t})^{(k)}\left(\mathbf{x}_t^{\alpha_n}-\boldsymbol{\mu}_{q,c}^{(k+1)}\right)\left(\mathbf{x}_t^{\alpha_n}-\boldsymbol{\mu}_{q,c}^{(k+1)}\right)^t}{\sum_t\sum_n(s_{t,q}r_{t,n})^{(k)}(z_{c,t})^{(k)}} \quad (40)$$

for all states (q) in the initial state space of size Q and c components in the mixture.

In order to speed up the method, the expected values in (22) and (26), defined in Section II, can be approximated as 0 or 1 using the Viterbi decoding algorithm (18). This is similar to the approach used for standard HMM in the segmental k-means approach for training [25].

IV. Local Time Warp for Dynamic Features

This section presents a method for modeling the dynamic temporal characteristics of speech. The method involves frame-specific warping of the time axis to facilitate locally optimum time resolution in the computation of the dynamic cepstrum. Temporal characteristics of the dynamic cepstrum are often described by way of the average modulation spectrum. However, characterizing the dynamic cepstrum have not previous approaches attempted to represent local temporal variability in speech. This section describes how the MATE framework can be applied to selecting the frame-specific temporal resolution for the dynamic features in MFCC feature analysis.

Temporal variability is implicitly modeled by the nonlinear time alignment performed by the Viterbi algorithm. However, this nonlinear time alignment does not extend to the computation of frame level features. This is because the length of the dynamic parameter analysis window remains static despite the fact that the rate of the audio events is not fixed [26].

Allowing local optimization of the temporal resolution over which the first- and second-order difference cepstra are computed is equivalent to a nonuniform sampling of the time scale for dynamic features. In the case of dynamic cepstrum computation, the temporal window length used for computing first- and second-order dynamic cepstrum coefficients is selected individually for each frame to produce more accurate dynamic features. In this case, Viterbi search is also modified so that an optimum time resolution can be chosen in the search process. For the purpose of computing the dynamic cepstrum, the frame update interval can be compressed or expanded by a "resolution factor" β to obtain a new update interval.

Given an ensemble of resolution factors $\mathcal{B} = \{\beta_n\}_{n=1}^N$, the dynamic features can be locally optimized using the modified Viterbi algorithm in (18), where the frequency-warped parameters $\mathbf{x}_t^{\alpha_n}$ are replaced by the new β-dependent parameters $\mathbf{x}_t^{\beta_n}$. The same local constraints that were applied in Section III to limit the rate of change of the frequency-warping parameters will also be applied to the temporal resolution parameters.

To account for this, we define $\tilde{\mathbf{o}}_{t'}$ as the log filter bank output vector computed for each sample. The original filter bank output vectors $\mathbf{o}_t$ are a downsampled version of $\tilde{\mathbf{o}}_{t'}$, where $t' = M\cdot t$, and M is the standard frame update interval (the sample to frame rate subsampling factor). In this resampling scenario, t is the frame index defined in Section II and t' is the time index for a sample rate of one frame per sample. A window of filter bank outputs centered at frame t can be written in terms of $\tilde{\mathbf{o}}_{t'}$ vectors as

$$\tilde{\mathbf{O}}_t = \left\{\tilde{\mathbf{o}}_{M(t-\lfloor L/2\rfloor)},\ldots,\tilde{\mathbf{o}}_{Mt},\ldots,\tilde{\mathbf{o}}_{M(t+\lfloor L/2\rfloor)}\right\} \quad (41)$$

where the matrix $\tilde{\mathbf{O}}_t$ is size $B \times ML+1$.

The following time projection matrix $\tilde{\mathbf{W}}_T^{\beta}$ can perform the local time warping of factor β since it is defined to obtain the dynamic features and the downsampling by a factor $\beta\cdot M$ simultaneously

$$\tilde{w}_{T,i',j}^{\beta} = \begin{cases} w_{T,i,j} & \text{if } i' = \lfloor i\cdot\beta M\rfloor, \\ 0 & \text{elsewhere.}\end{cases} \quad (42)$$

In (42), the values not equal to zero apply the $\beta\cdot M$ downsampling and the time projection, since $w_{T,i,j}$ are the components of the previously defined standard time projection matrix $\mathbf{W}_T$ $(L\times L')$, with $i = 1,\ldots,L$ and $j = 1,\ldots,L'$.

Using notation similar to that used to describe localized frequency warping in (6), the feature vectors obtained from localized time warping can be written as

$$\mathbf{x}_t^{\beta} = \mathrm{vec}\left(\mathbf{W}_F \cdot \tilde{\mathbf{O}}_t \cdot \tilde{\mathbf{W}}_T^{\beta}\right) \tag{43}$$

where the time projection matrix is the previously defined in (42).

By estimating β using the augmented state space decoding algorithm in Section III, frame-specific temporal resolution is obtained for dynamic features that maximizes the likelihood of the utterance with respect to the original HMM. In the search of the optimum time factors, we also apply the continuity constraints described in Section III-B.

V. Experiments

Three sets of experiments have been performed. In the first one, a general study of the MATE framework in connected digits for speaker-independent ASR has been performed. In the second, one a study of speaker variability was performed, testing the ability of the model to compensate for severe inter-speaker mismatch. Finally, in the third set of experiments, the MATE-based speaker normalization techniques were evaluated on a speech corpus containing speech from a population of children diagnosed with speech disorders. The task domains were based on a Spanish language subset of the Speech-Dat-Car database [27], the "A" subset of the Aurora 2 task [28], which includes subway, babble, car, and exhibition noises for the first experiment, the TIDIGITS database [29] for the second experiment, and a Spanish language corpus of speakers with impaired and unimpaired speech [13], [14], which will be analyzed in the third experiment. The experiments relating to warping of the frequency axis in the MATE framework were performed using the approximation to the Jacobian term described in Appendix I.

A. Speaker-Independent Models

For speaker-independent experiments, HMM word models with 16 states and three Gaussians per state were used to model the vocabulary of spoken digits. Initial and final silence were modeled by three-state HMMs with six Gaussians per state. Inter-word silence was modeled by one state HMM with six Gaussians. The parameters used in the experiments were the standard [30] and the advanced ETSI front end [31], both with a window size of 25 ms and a 10-ms frame update interval. Twelve cepstrum coefficients and the energy are the static feature vector. Then, velocity and acceleration parameters were computed for a window of nine static frames after a time projection with the $\mathbf{W}_T$ matrix of size 9×3 as described in Section II, resulting in a total of 39 parameters. The baseline models were obtained with 20 training iterations and used to build initial MATE models according to the observation densities given in (8). The number of transformations in all the experiments for the local warping factors was set up to $N = 5$ for time and frequency MATE.

Retrained MATE models were obtained after one iteration by using the training formulas in Section III. Viterbi alignment for the expected values [25] was used, since in the experiments, we did not experience any significant difference by training after computing the expected values from the augmented forward and backward auxiliary functions $a_{q,n}(t)$ and $b_{q,n}(t)$ compared to segmental k-means. This result is similar to what many authors have found comparing standard HMM trained with EM or segmental k-means.

TABLE I
Clean Test Recognition Results in WER%

DataBase	Aurora2	Speech-Dat-Car
baseline	0.90	0.88
VTLN	0.85 (+6%)	0.81 (+8%)
retrained-VTLN	0.87 (+3%)	0.77 (+13%)
MATE-freq	0.75 (+17%)	0.68 (+23%)
retrained-MATE-freq	0.74 (+17%)	0.67 (+24%)
MATE-time	0.88 (+2%)	0.75 (+15%)

Table I shows the experimental results obtained using noise-free speech in order to compare performance of the local temporal and frequency optimization techniques to the original baseline performance using the standard ETSI front end. The experiments were performed on two data sets. First, the *Aurora2* clean training set was used which consists of 8440 phrases (27 727 digits) and test set consists of 4004 phrases (13 159 digits). Second, a subset of Spanish language Speech-Dat-Car was used. A noise-free close-talk microphone training set consisting of 1896 phrases (9948 digits) and test set consisting of 1087 phrases (5701 digits) were used.

The results in Table I show that the MATE-frequency model reduces WER with respect to the baseline system to VTLN for both task domains. It is also clear from Table I that the temporal optimization performed using the MATE decoder (MATE-time) also reduced WER with respect to the baseline with a 15% reduction obtained for the Speech-Dat-Car corpus. This improvement may be a result of the greater temporal variability of the speech in this corpus since it was collected under a variety of driving conditions. Nevertheless, MATE-time retraining did not provide any further improvement. Additional experiments should be performed to conclude the effect of MATE-time on a spontaneous speech corpus that exhibited a wider variety of speaking styles. The best results were obtained by using retrained-MATE-frequency models, given a WER improvement of 23% for Aurora 2 and 24% for the Speech-Dat-Car subset. In these experiments, a maximum of 20% frequency warping for Speech-Dat-Car and 15% for Aurora2 (i.e., α between 0.8 and 1.2 and between 0.85 and 1.15, respectively) and a 10% of temporal resolution deviation (i.e., β between 0.9 and 1.1) were allowed. N was five warpings for all the cases. The best configuration for VTLN was used in this comparison, which is the same as in [2], α ranging from 0.88 to 1.22 and $N = 11$ warpings. VTLN optimal range is narrower than MATE since VTLN extracts the best average warping factor, and MATE is a frame-by-frame estimation.

The best values of improvement of MATE in Table I for both databases have been evaluated for statistical confidence with McNemar's test [32], since the test database for baseline and MATE is the same. This showed Aurora2 results to be significant within the confidence interval of 95%, but they were not for

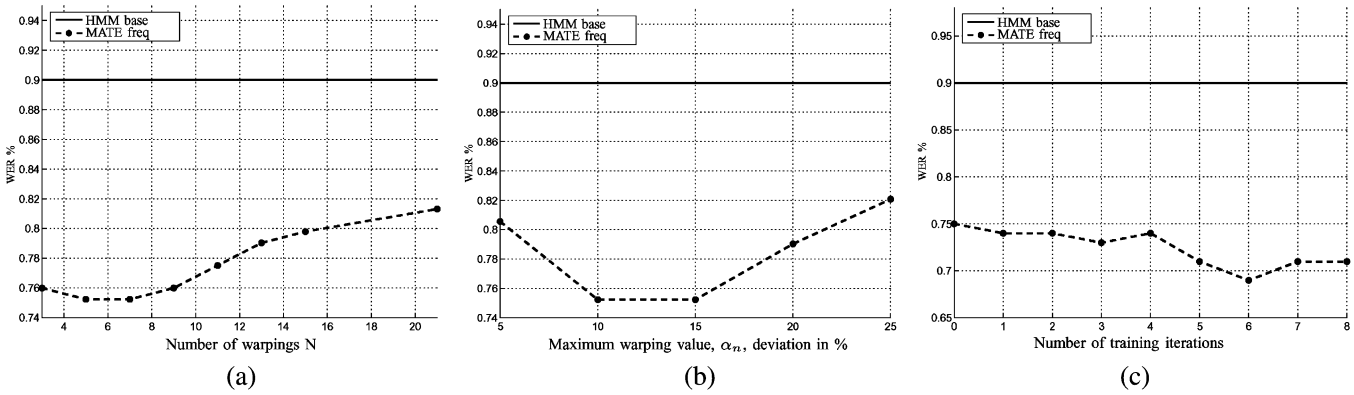

Fig. 6. Results for some values of N and maximum α on Aurora2 clean test for MATE freq. with respect to the baseline. (a) Error rate of MATE freq. for different number of warpings N and maximum warping, $\alpha_n = 10\%$. (b) Maximum warping value α_n deviation in % with $N = 5$. (c) Number of iterations in training, where 0 means no retrained model, with $\alpha_n = 15\%$ and $N = 5$.

the Speech-Dat-Car in this interval, due to the size of the database. Results from Speech-Dat-Car were statistically significant within the confidence interval of 90%.

Fig. 6(a) and (b) shows the performance of MATE-freq with respect to the number of frames N and the maximum α factor in WER%. It can be seen that there is an optimum value, but MATE improves the baseline for all the common values. Fig. 6(c) shows a slight improvement of the performance of MATE-freq when retraining with respect to the number of iterations.

In this database, an experiment on cepstrum resolution was performed. The dimensionality of the standard cepstrum feature vector for ASR of 12 coefficients has been empirically derived to provide a minimum WER in most tasks. As this dimensionality grows, lower results are achieved due to the fact that higher resolution spectral estimations begin to capture too much speaker-dependent variability. In this experiment, it is shown that thanks to the local warpings, an increase in the dimensionality of the frequency features still provides improvements in results since variability is partially reduced.

The experiment was performed with speech-Dat-Car and Aurora2 clean test databases, 16 state word models with one Gaussian component and the same silence models as previous experiments. Different values for the number of cepstrum coefficients were tested while the number of derivatives was fixed as 2. In Fig. 7, it is shown that, as expected, the best baseline result is obtained for 12 cepstrum coefficients. The results obtained for MATE with local frequency normalization, with α ranging from 0.8 to 1.2 and $N = 5$ warpings show a different tendency, the WER still decreases for a larger number of cepstrum coefficients. The minimum error for MATE was found for 16 cepstrum coefficients.

These results are very interesting since they show the ability of MATE to reduce local variability by aligning at frame level for frequency axis. If the acoustic events in the short time spectrograms are aligned, then the variability is partially removed so that events of the same nature are compared. The effect of small unalignment is more noticeable for small-scale acoustic events. Consequently, if the spectrograms are aligned, more cepstrum (DCT) coefficients can be used in order to capture small scale spectrogram details and, eventually, reduce the WER. These experiments show that local alignments are useful to improve the front-end, because the discriminant speech events can be captured with more detail. Even if the relative gain with respect to the system with 12 coefficients is not statistically significant, further attention would be deserved for this trend.

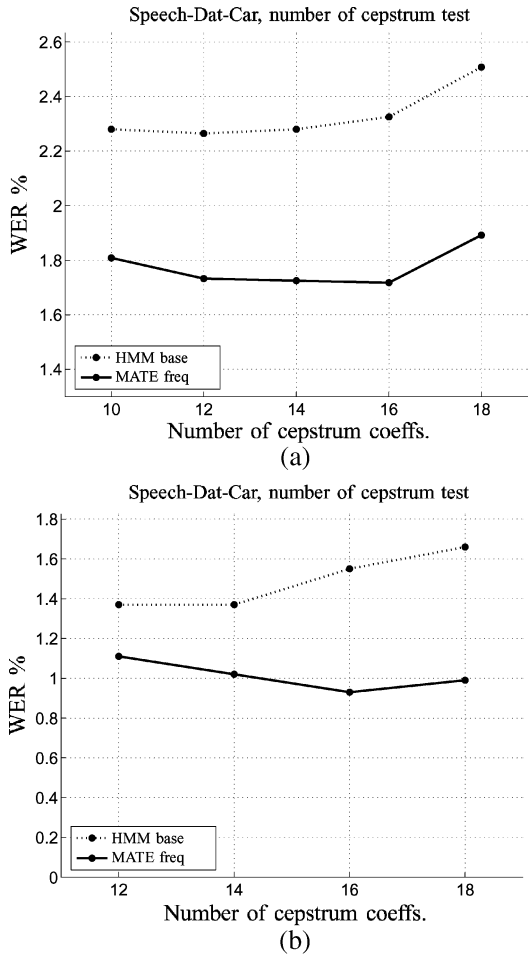

Fig. 7. Mean WER for baseline system and MATE for different number of cepstrum coefficients for one Gaussian per state digit word models. (a) Aurora2 database.(b) Speech-Dat-Car database.

This effect can be argued as follows: let us assume that there exists a small aligning noise compared with the size of the bigger structures we are measuring, such as formants. Then, the higher order cepstrum projections would be the most affected

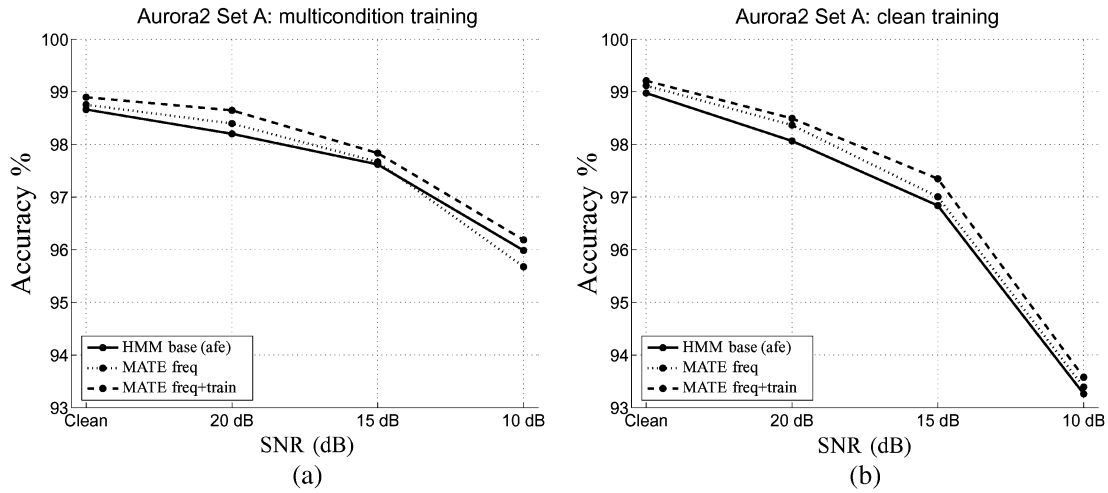

Fig. 8. Aurora2 set "A" average results for the Advanced ETSI Front-End. (a) Multicondition training. (b) Clean training, both from clean signal to 10 dB.

by an unalignment of small scale, compared to the first cepstrum vectors, which tend to capture bigger structures and remain unaffected by a small unalignment. This is clear if we think that keeping the first vectors of a DCT projection is equivalent to smoothing the spectrogram. An eventual alignment can be of benefit in terms of sharpness in the pattern we are about to learn. Under this reasoning, Fig. 7 results demonstrate that the alignment is being effectively produced, since an increase in the number of projection vectors reduces the WER, while this is not happening in the baseline HMM case. Similarly to the dynamic programming alignment effect for speech utterances aligned with audio templates or models, this local alignment helps the classifier to compare local events of the same nature together.

Although the main motivation of the paper is not a noise robustness study, an experiment was performed in order to evaluate the degradation of the framework benefits under noise conditions. Fig. 8 shows the results of experiments performed using Aurora 2 test set "A" noise conditions. Clean and multicondition training cases were evaluated. In order to increase the noise robustness of the model, a modification of Advanced ETSI Front-End, AFE, parameters for frequency warping, performed as in (6), were used. The average improvement of MATE, with α ranging from 0.8 to 1.2 and $N = 5$ warpings, compared to unmodified Advanced ETSI front end features found for the complete test set "A," was 6.77% for the multicondition training and 9.64% for clean training. This corresponds to an average improvement of 8.20% for all the noise conditions. Greater improvements are obtained in signals over 10 dB of SNR, as shown in Fig. 8. These results show that the use of Advanced ETSI front end jointly with the MATE-frequency decoder gives similar WER improvements to those obtained with clean speech in Table I.

B. Speaker Variability

An experiment investigating inter-speaker variability was performed in order to evaluate the performance of the new models under mismatched conditions. Multiple experiments have been carried out combining different speaker group training and testing conditions as is shown in this section. In a previous study, extensive experiments were performed focused on recognition on speech from children with the TIDIGITS corpus [33]. In this work, the task domain for the mismatch conditions was also the TIDIGITS corpus. This 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). In this experimental study, seven partitions were defined in the training and testing sets: "boy," "girl," "man," "woman," "boy + girl," "man + woman," and "all." For all those partitions, word models consisting of 16 HMM states per word and a begin-end silence model containing three states were trained. The standard ETSI feature set plus energy and their first and second derivatives, as defined in Section II, were used. In this study, the speaker variability reduction on a high mismatch task was evaluated. This experiment was performed on a subset of the TIDIGITS corpus containing only isolated digits in order to test the ability of the proposed method to reduce inter-speaker mismatch when only limited training data are available (3586 isolated digit utterances for training and 3582 for testing). As mismatch conditions were tested, models for MATE were not reestimated and the simple expansion of states was performed, a 20% of deviation for α was set, i.e., α between 0.8 and 1.2, and $N = 5$. The results of this study are shown in Fig. 9, the horizontal axis of this figure represents the number of Gaussian mixtures per state in the HMM. HMM models were trained from utterances taken from each of the above seven data partitions. For each case, the models were tested with data from of all the partitions excluding matching cases (i.e., training model from the "boy" labeled partition and testing with "man" labeled partition), giving a total of 42 mismatch testing experiments. The WER in Fig. 9 was calculated as the average of the WER of all those experiments. It is clear from Fig. 9 that the effect of overtraining is observable as the number of Gaussian components grow, since it is a small data set. However MATE can reduce the WER effectively in this kind of situation.

Comparing the best obtained result with respect to the HMM baseline with four Gaussians, the mean WER reduction is a 79% for MATE and a 14% of reduction for VTLN in this same situation. In this experiment, the standard deviations of the WER values were 6.6% for the baseline, 1.5% for MATE and 5.7%

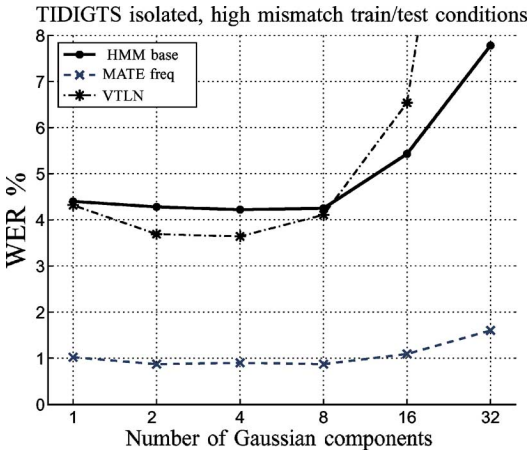

Fig. 9. Mean WER in the speaker train and test models mismatch isolated digits experiment (from TIDIGITS corpus) for the baseline and MATE.

for VTLN. Two reasons can be given to explain the results obtained for VTLN. First, since isolated digit utterances are very short, then when an error is produced for the hypothesis, the optimization of the warping factor α is worse than in longer utterances. In longer utterances, most of the words of the hypothesis could be correct, then the factor is better estimated by maximum likelihood and part of the errors could be recovered in the second pass. The worst are the hypotheses, the higher is WER after VTLN. At a certain level of errors, VTLN even degrades the performance of the system. Second, in on order to compare the effect of the local warping, the number of transformations in VTLN has been fixed to be the same as in MATE, in this experiment $N = 5$, which is a much rougher step than the usual ones for VTLN (i.e., $N = 11$ or 13).

This is an artificial set, but there are many applications were there exists train and test mismatch. These experiments show the ability of generalization of MATE in controlled speaker mismatch situations where unseen data are recognized (e.g., test children with trained adult models), since the mechanism of adaptation to the speaker lies in the model and then less training data are required. This can beusefull for many ASR systems where there is no prior information of the kind of speaker.

C. A Speech Corpus From Impaired and Unimpaired Speakers

Speech disorders such as dysarthria, dyslalia, dysglossia, or aphasia [34] dramatically affect the communication abilities of those who suffer them [35]. Although the range of causes and symptoms of these disorders is very wide, these speech handicaps can mainly be originated by one of these three reasons. First, different forms of brain damage, like cerebral palsy, or stroke can lead to dysarthria or aphasia [34], with limited control of the organs used for speech production. Second, when any of the organs of the articulatory system (tongue, mouth, vocal chords, etc.) is affected in its morphology or movement, it may lose its ability to generate correct speech. This situation leads to the presence of a dysglossia. Finally, when there is no physical disability that affects speech, and usually related to a type of developmental disabilities like Down Syndrome, a dyslalia can appear as the other main type of speech pathology. In this situation, the patient mistakes or misses different sounds and phonemes during the production of the speech. The study of all of these disorders from a phonological point of view has clearly shown how they affect the normal production of speech, resulting in a sometimes critical variation of the main parameters of speech [36].

The corpus used for this work was recorded by the Department of Signals and Communications from the University of Las Palmas de Gran Canaria (Spain). The corpus contains 1077 utterances of unimpaired speech and 2470 utterances of impaired speech recorded by several speakers by terms of age and gender. The phonological content of the corpus consists of utterances of the 57 words from the "Induced Phonological Register." This set of words contains a phonetically rich selection of words that includes nearly all the phonemes in the Spanish language, as well as diphthongs and other singularities of the language. The length of the words is balanced and range from one-syllable words to four-syllable words. The corpus was originally used for the research in identification and classification of impaired speech [14] and also was evaluated in [13].

When dealing with impaired speech, the selection of the HMM structure does not necessarily have to agree with the traditional structures [37]. The experiments performed evaluate a number of different HMM word models with increasing state number. In Fig. 10, it can be seen that the optimal size of the word-based HMM was around 24 states per model. Given the small amount of data available, the number of Gaussians in the HMM was chosen to be one Gaussian component per state, since more components increased the WER as result of overtraining. The baseline in the three cases of matched-model for unimpaired speech and impaired speech is shown in Fig. 10. The results over the four train-test sets present significantly higher WER for the impaired speech, showing the difficulty of this ASR task. The results of the frequency MATE with α between 0.8 and 1.2 are also shown. There is a noticeable reduction in the WER, 11.14% for impaired speech in matched train and test conditions [Fig. 10(b)] and 38.81% for unimpaired speech in matched train and test conditions [Fig. 10(a)] and 10.63% in mismatch conditions [Fig. 10(c)] i.e., train data is unimpaired speech and test data is impaired speech. We see that the improvement is smaller when evaluated on the impaired speech corpus. The reason for this is due to the fact that there are a large number of disfluencies and severe mispronunciations in this corpus. The large number of recognition errors that arise from these effects represents a larger portion of the overall error rate than any of the other speech corpora used in this study. Clearly, it is necessary to use other means for modeling this class of errors. These results demonstrate that the impact of the frequency variability on the ASR performance can be reduced by more complex models in which variability is taken into account.

VI. Discussion and Conclusion

An augmented state space acoustic decoding method for speech variability normalization (MATE) has been presented. The algorithm can be considered to be an extension to HMM and VTLN methods. The technique provides a mechanism for either the spectral warping or the dynamic feature computation to be locally optimized. In the MATE decoder, the optimum sequence of states in the augmented state space is identified using a modified Viterbi algorithm. It allows frame specific

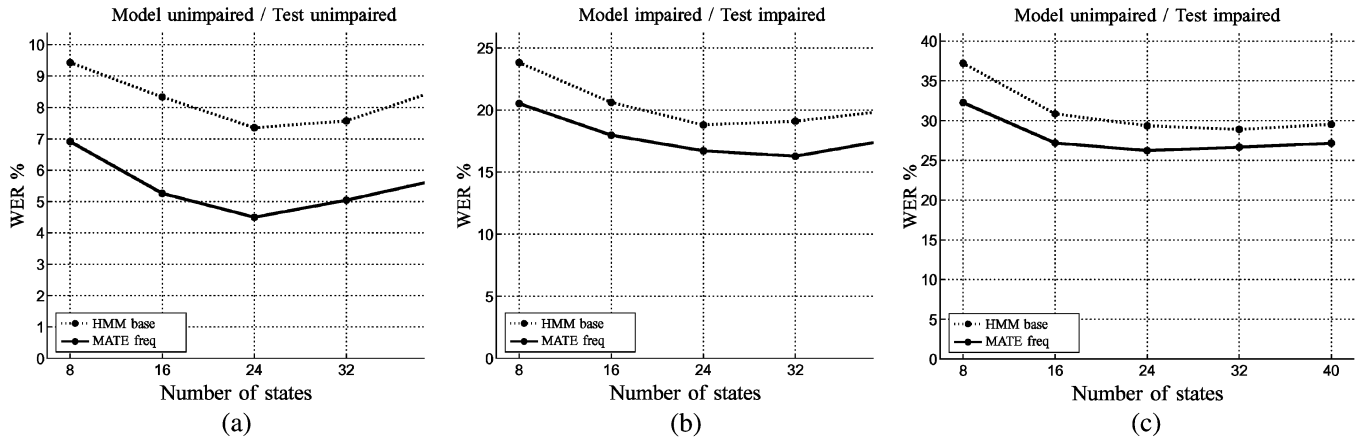

Fig. 10. Mean WER for baseline system and MATE for different number of HMM states on impaired speech. (a) Unimpaired speech is used for train and test the models. (b) Models are trained with unimpaired speech and tested with impaired speech. (c) Models are trained and tested with impaired speech.

spectral warping functions or temporal resolution for the dynamic features to be estimated as part of search. It includes frame-specific transformations of the speech, by means of expansion of the HMM state space while at the same time not increasing substantially the number of parameters used in the model or the computational complexity during decoding.

The MATE approach was evaluated on four speech corpora including Aurora 2, the Spanish Speech-Dat-Car, and the TIDIGITS, and a Spanish language corpus of speakers with impaired and unimpaired childrens' speech. Experiments were performed under clean and noisy speech conditions. The method has been compared in the context of existing methods for frequency warping-based speaker normalization and existing methods for computation of dynamic features. The results have shown that MATE is an effective method for compensating the local variability of the speech compared to VTLN methods. MATE-frequency gives more than a factor of two in reduction of WER over that obtained using VTLN. MATE-time was shown to provide more than 15% WER reduction when used in real and stressed situations.

These results suggest that there may be other speech transformations that may be applied in the context of the constrained search algorithm described here. The MATE decoder can be seen as a method for making HMMs more "locally elastic" and providing a mechanism for better tracking the dynamics of speech variability. Application to domains that can be characterized as having more extreme speaker variability, arising from increased vocal effort or task related stress, will provide a further indication of the potential benefits of these techniques. In [38], existing VTLN methods are studied, and it is suggested how effects of stress on the vocal tract are not uniform during an utterance. The proposed method's main motivation is to model these kinds of situations.

The experiments related to severe speaker mismatch reduction tested on TIDIGITS corpus have shown significant WER reductions. When the system was tested with speakers from different populations than those whose data were used to train the models, the improvement reached up to 79% for the best baseline system case, in a relatively small subset of the corpus.

Also we have presented experiments of speech uttered by people with several kinds of speech disorders. It has been shown how speech disorders affect the speech signal when comparing impaired and unimpaired ASR results. The final recognition results show improvements up to 17% in WER, which can be of help in computer-aided systems. These results point out that a warping method like MATE can reduce variations in speech for ASR.

Clearly, the computational load of the MATE decoder grows with the number of transformations N. According to the constraint (20), the number of transitions from an expanded state grows a factor of 3, then the total complexity is increased by a factor of $3N$. This complexity could be reduced by more efficient pruning algorithms.

For future work, we should consider the possibility of reducing the impact of time and frequency distortions simultaneously in a time–frequency warping method. Better results should be expected by taking into account interaction between both domains.

APPENDIX I

The likelihood normalizing term, the determinant of the Jacobian, must be present when a pdf is defined after a transformation or function of a given variable, as in (8). The warped static feature vector $\mathbf{c}_t^{\alpha_n}$ can be expressed as a linear function of the unwarped vector $\mathbf{c}_t$ (5), as in [39]. Then we express the warped dynamic feature vector $\mathbf{x}_t^{\alpha_n}$ as

$$\mathbf{x}_t^{\alpha_n} = f_{\mathbf{r}_t}(\mathbf{x}_t) \simeq \tilde{\mathbf{A}}^{\alpha_n} \cdot \mathbf{x}_t \tag{I-1}$$

where $\mathbf{x}_t^{\alpha_n}$ is the dynamic feature vector at time t warped with the transformation indicated by $\mathbf{r}_t$, i.e., the warping factor α_n for the component $r_{t,n} = 1$ and zeros elsewhere.

The dynamic feature vector transformation matrix $\tilde{\mathbf{A}}^{\alpha_n}$ is now defined for a particular case, the ETSI standard front-end. In ETSI standard, the static feature vector has $C = 12$ cepstrum components and the log energy. The number of dynamic streams for the dynamic projection, L' in (3), is $L' = 3$, i.e., static, first, and second time derivatives. The matrix $\tilde{\mathbf{A}}^{\alpha_n}$ is then

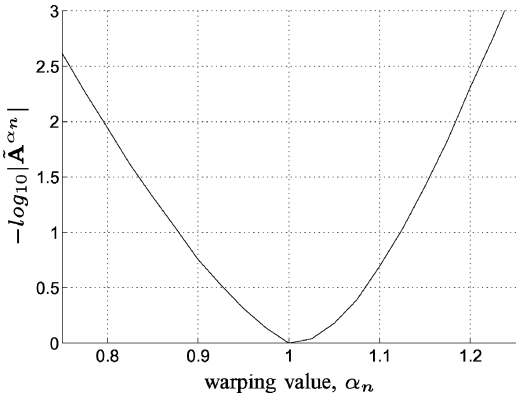

Fig. 11. Plot of $-log_{10}|\tilde{\mathbf{A}}^{\alpha_n}|$ for 39-dimensional ETSI cepstral coefficients as a function of α_n.

a three block diagonal matrix of size 39 × 39 for dynamic feature vector warping

$$\tilde{\mathbf{A}}^{\alpha_n} = \begin{pmatrix} \mathbf{A}^{\alpha_n} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}^{\alpha_n} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}^{\alpha_n} \end{pmatrix} \tag{I-2}$$

being A^{α_n} the warping matrix for the static features, (of size 13 × 13) as defined in (5).

After (I-1), the Jacobian determinant can be calculated as

$$J(n) = \left| \frac{\delta f_{\mathbf{r}_t}(\mathbf{x}_t)}{\delta \mathbf{x}_t} \right| = |\tilde{\mathbf{A}}^{\alpha_n}| = |\mathbf{A}^{\alpha_n}|^3 \tag{I-3}$$

being one the nth component of $\mathbf{r}_t$, $r_{t,n} = 1$, and zeros elsewhere. Where $\mathbf{A}^{\alpha_n}$ is the static cepstrum transformation matrix.

This matrix can be obtained analytically given the front-end definition [4], or in a constrained MLLR model-based estimation [40], or simply as a simple linear regression as in [12]. In the case of linear regression, we can use the well-known result of the multidimensional regression minimum square error (MSE) to estimate the best regression matrix between the warped cepstrum obtained from (4) and the unwarped cepstrum [12]. We have depicted the values of the determinant of the empirically calculated matrices for the ETSI front-end with 39 dimensions in Fig. 11.

In the experimental section, the results related with MATE were calculated with the approximation of $J(n) \simeq 1$, since we reported similar performance in our experiments including or approximating the Jacobian term [12].

APPENDIX II

The expanded auxiliary functions, $a_{q,n}(t)$ and $b_{q,n}(t)$, are defined as follows

$$a_{q,n}(t) = p\left(\{\mathbf{x}_t\}_1^t, s_{t,q} = 1, r_{t,n} = 1\right) \tag{II-1}$$
$$b_{q,n}(t) = p\left(\{\mathbf{x}_t\}_{t+1}^T, s_{t,q} = 1, r_{t,n} = 1\right). \tag{II-2}$$

We can express (II-1) and (II-2) as result a marginalization over all possible values of n' and q' as

$$a_{q,n}(t) = \sum_{q',n'} p\left(\{\mathbf{x}_t\}_1^t, s_{t,q} = 1, r_{t,n} = 1 \right. \\ \left. s_{t-1,q'} = 1, r_{t-1,n'} = 1\right) \tag{II-3}$$
$$b_{q,n}(t) = \sum_{q',n'} p\left(\{\mathbf{x}_t\}_{t+1}^T, s_{t,q} = 1, r_{t,n} = 1 \right. \\ \left. s_{t+1,q'} = 1, r_{t+1,n'} = 1\right). \tag{II-4}$$

Then, applying the first-order HMM assumptions, as in (13), and splitting the sequence and applying Bayes rule follows:

$$a_{q,n}(t) = \sum_{q',n'} \left[p\left(\{\mathbf{x}_t\}_1^{t-1}, s_{t-1,q'} = 1, r_{t-1,n'} = 1\right) \right. \\ \cdot p(s_{t,q} = 1, r_{t,n} = 1 | s_{t-1,q'} = 1, r_{t-1,n'} = 1) \\ \left. \cdot p(\mathbf{x}_t | s_{t,q} = 1, r_{t,n} = 1)\right] \tag{II-5}$$
$$b_{q,n}(t) = \sum_{q',n'} \left[p\left(\{\mathbf{x}_t\}_{t+1}^T, s_{t+1,q'} = 1, r_{t+1,n'} = 1\right) \right. \\ \cdot p(s_{t,q} = 1, r_{t,n} = 1 | s_{t+1,q'} = 1, r_{t+1,n'} = 1) \\ \left. \cdot p(\mathbf{x}_{t+1} | s_{t+1,q'} = 1, r_{t+1,n'} = 1)\right] \tag{II-6}$$

where we can identify the first term as the recursive expression as defined in (II-1) and (II-2), the second term a transition probability, and the third the pdf of the augmented state space as defined in (8). Then, the recursive expressions are

$$a_{q,n}(t) = \sum_{q',n'} a_{q',n'}(t-1) \cdot \pi_{q',n',q,n} \\ \cdot p\left(\mathbf{x}_t^{\alpha_n} | s_{t,q} = 1\right) \cdot J(n) \tag{II-7}$$
$$b_{q,n}(t) = \sum_{q',n'} b_{q',n'}(t+1) \cdot \pi_{q',n',q,n} \\ \cdot p\left(\mathbf{x}_{t+1}^{\alpha_n} | s_{t+1,q'} = 1\right) \cdot J(n'). \tag{II-8}$$

Then, the previous expressions for the expected values (25) and (31) can be calculated using these auxiliary functions as

$$(s_{t,q} r_{t,n})^{(k)} \tag{II-9}$$
$$= \left[p\left(\{\mathbf{x}_t\}_1^t, s_{t,q} = 1, r_{t,n} = 1 | \boldsymbol{\Theta}^{(k)}\right)\right. \tag{II-10}$$
$$\left. \cdot p\left(\{\mathbf{x}_t\}_{t+1}^t, s_{t,q} = 1, r_{t,n} = 1 | \boldsymbol{\Theta}^{(k)}\right)\right] \\ \cdot \frac{1}{p\left(\mathbf{X} | \boldsymbol{\Theta}^{(k)}\right)} \tag{II-11}$$
$$= \frac{a_{q,n}^{(k)}(t) b_{q,n}^{(k)}(t)}{\sum_{q,n} a_{q,n}^{(k)}(T)} \tag{II-12}$$

and

$$\left(s_{t-1,q}r_{t-1,n}s_{t,q'}r_{t,n'}\right)^{(k)} \tag{II-13}$$

$$= \Big[p\left(\{\mathbf{x}_t\}_1^{t-1}, s_{t-1,q}=1, r_{t-1,n}=1|\boldsymbol{\Theta}^{(k)}\right) \tag{II-14}$$

$$\cdot p\left(\{\mathbf{x}_t\}_{t+1}^{t}, s_{t,q'}=1, r_{t,n'}=1|\boldsymbol{\Theta}^{(k)}\right) \tag{II-15}$$

$$\cdot p\left(s_{t,q'}=1, r_{t,n'}=1|s_{t-1,q}=1, r_{t-1,n}=1, \boldsymbol{\Theta}^{(k)}\right) \tag{II-16}$$

$$\cdot p\left(\mathbf{x}_t|s_{t,q'}=1, r_{t,n'}=1, \boldsymbol{\Theta}^{(k)}\right)\Big] \cdot \frac{1}{p\left(\mathbf{X}|\boldsymbol{\Theta}^{(k)}\right)} \tag{II-17}$$

$$= \frac{a_{q,n}(t-1)\cdot b_{q',n'}(t)\cdot \pi_{q,n,q',n'}\cdot p\left(\mathbf{x}_t^{\alpha_n}|s_{t,q'}=1, \boldsymbol{\Theta}^{(k)}\right)\cdot J(n')}{\sum_{q,n} a_{q,n}(T)}. \tag{II-18}$$

REFERENCES

[1] A. Andreou, T. Kamm, and J. Cohen, "Experiments in vocal tract normalization," in *Proc. CAIP Workshop: Frontiers in Speech Recognition II*, 1994, CD-ROM.

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

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

[4] M. Pitz and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 930–944, Sep. 2005.

[5] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Comput. Speech Lang.*, vol. 12, pp. 75–98, 1998.

[6] K. Weber, "HMM mixtures (HMM2) for robust speech recognition," Ph.D. dissertation, Swiss Federal Inst. of Technol. Lausanne (EPFL), Lausanne, Switzerland, 2003.

[7] G. Gravier, "Analyse statistique á deux dimensions pour la modélisation segmentale du signal de parole—application á la reconnaissance," Ph.D. dissertation, Ecole Nationale Supérieure des Télécommunications (ENST), Paris, France, Jan. 2000.

[8] T. Fukada and Y. Sagisaka, "Speaker normalized acoustic modeling based on 3-D Viterbi decoding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Seattle, WA, 1998, vol. 1, pp. 437–440.

[9] A. Miguel, R. Rose, E. Lleida, L. Buera, A. Ortega, and O. Saz, "Decodificador Eficiente para Normalización del Tracto Vocal en Reconocimiento Automático del Habla en Tiempo Real," in *Proc. III Jornadas sobre Tecnologías del Habla*, Valencia, Spain, 2004, pp. 269–274.

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

[11] R. Rose, A. Keyvani, and A. Miguel, "On the interaction between speaker normalization, environment compensation, and discriminant feature space transformations," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Toulouse, France, May 2006, pp. 985–988.

[12] A. Miguel, E. Lleida, A. Juan, L. Buera, A. Ortega, and O. Saz, "Local transformation models for speech recognition," in *Proc. Interspeech—Int. Conf. Spoken Lang. Process.*, Pittsburgh, PA, Sep. 2006, pp. 1598–1601.

[13] O. Saz, A. Miguel, E. Lleida, A. Ortega, and L. Buera, "Study of time and frequency variability in pathological speech and error reduction methods for automatic speech recognition," in *Proc. Interspeech—Int. Conf. Spoken Lang. Process.*, Pittsburgh, PA, Sep. 2006, pp. 993–996.

[14] J. L. Navarro-Mesa, P. Quintana-Morales, I. Pérez-Castellano, and J. Espinosa-Yañez, "Oral Corpus of the Project HACRO (Help Tool for the Confidence of Oral Utterances)," Dept. Signal Commun., Univ. of Las Palmas de Gran Canaria, Spain, May 2005.

[15] A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Albuquerque, NM, Apr. 1990, pp. 845–848.

[16] M. J. F. Gales and S. Young, "An improved approach to the hidden Markov model decomposition of speech and noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, San Francisco, CA, Mar. 1992, pp. 233–236.

[17] E. Lleida and R. Rose, "Utterance verification in continuous speech recognition: Decoding and training procedures," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 2, pp. 126–139, Mar. 2000.

[18] G. Zweig and S. Russell, "Probabilistic modeling with Bayesian networks for automatic speech recognition," in *Proc. Int. Conf. Spoken Lang. Process.*, Sydney, Australia, Nov. 1998, pp. 3010–3013.

[19] S. Furui, "Speaker independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 1, pp. 52–59, Feb. 1986.

[20] A. Juan and E. Vidal, "On the use of Bernoulli mixture models for text classification," *Pattern Recognition*, vol. 35, no. 12, pp. 2705–2710, Dec. 2002.

[21] M. Pitz and H. Ney, "Vocal tract normalization as linear transformation of MFCC," in *Proc. Eur. Conf. Speech Commun. Technol.*, Geneva, Switzerland, Sep. 2003, pp. 1445–1448.

[22] Z. Ghahramani, "Learning dynamic bayesian network," in *Lecture Notes In Computer Science*. New York: Springer, 1997, vol. 1387, pp. 168–197.

[23] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the *EM* algorithm," *J. R. Statist. Soc.*, vol. 39, no. 1, pp. 1–21, 1977.

[24] L. R. Rabiner, *A Tutorial on HMM and Selected Applications in Speech Recognition*. San Mateo, CA: Morgan Kaufmann, 1988, ch. 6.1, pp. 267–295.

[25] B. H. Juang and L. R. Rabiner, "The segmental k-means algorithm for estimating the parameters of hidden Markov models," *IEEE Trans. Accoust., Speech, Signal Process.*, vol. 38, no. 9, pp. 1639–1641, Sep. 1990.

[26] J. Hant and A. Alwan, "A psychoacoustic-masking model to predict the perception of speech-like stimuli in noise," *Speech Commun.*, vol. 40, no. 3, pp. 291–313, May 2003.

[27] H. van den Heuvel, J. Boudy, R. Comeyne, S. Euler, A. Moreno, and G. Richard, "The Speechdat-Car multilingual speech databases for in-car applications: Some first validation results," in *Proc. Eur. Conf. Speech Technol.*, Budapest, Hungary, Sep. 1999, pp. 2279–2282.

[28] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ISCA ITRW ASR2000 Automatic Speech Recognition: Challenges for the Next Millennium*, Paris, France, Sep. 2000, pp. 18–20.

[29] R. G. Leonard, "A database for speaker-independent digit recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, San Diego, CA, Mar. 1984, pp. 328–331.

[30] "ETSI ES 202 050 v1.1.1 distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," Jul. 2002.

[31] "ETSI ES 201 108 v1.1.2 Distributed speech recognition; front-end feature extraction algorithm; compression algorithms," April 2000.

[32] L. Gillick and S. J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 1989, pp. 532–535.

[33] A. Potamianos and S. Narayanan, "Robust recognition of children's speech," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 603–616, Nov. 2003.

[34] F. L. Darley, A. E. Aronson, and J. R. Brown, "Differential diagnostic patterns of dysarthria," *J. Speech Hearing Res.*, vol. 12, no. 2, pp. 246–269, 1969.

[35] P. Green, J. Carmichael, A. Hatzis, P. Enderby, M. Hawley, and M. Parker, "Automatic speech recognition with sparse training data for dysarthric speakers," in *Proc. Eur. Conf. Speech Technol.*, Geneva, Switzerland, Sep. 2003, pp. 1189–1192.

[36] K. Croot, "An acoustic analysis of vowel production across tasks in a case of non fluent progressive aphasia," in *Proc. Int. Conf. Spoken Lang. Process.*, Sydney, Australia, Dec. 1998, pp. 907–910.

[37] J. R. Deller, D. Hsu, and L. J. Ferrier, "On the use of Hidden Markov Modelling for recognition of dysarthric speech," *Comput. Meth. Programs Biomed.*, vol. 35, no. 125, pp. 125–139, 1991.

[38] J. Hansen, "Morphological constrained feature enhancement with adaptive cepstral compensation (MCE-ACC) for speech recognition in noise and Lombard effect," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 598–614, Oct. 1994.

[39] M. Pitz, S. Molau, R. Schluter, and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," in *Proc. Eur. Conf. Speech Technol.*, Aalborg, Denmark, 2001, pp. 2653–2656.
[40] D. Kim, S. Umesh, M. J. Gales, T. Hain, and P. Woodland, "Using VTLN for broadcast news transcription," in *Proc. Int. Conf. Spoken Lang. Process.*, Jeju Island, Korea, Oct. 2004, pp. 1953–1956.

Antonio Miguel was born in Zaragoza, Spain, in 1977. He received the M.Sc. degree in telecommunication engineering from the University of Zaragoza (UZ), Zaragoza, Spain, in 2001. He is currently working towards the Ph.D. degree at UZ.

From 2000 to 2007, he was with the Communications Technologies Group, Department of Electronic Engineering and Communications, UZ, under a research grant. Since 2006, he has been an Assistant Professor in the same department. Currently, his research interest lies in the field of acoustic modeling for ASR.

Eduardo Lleida (M'89) was born in Spain in 1961. He received the M.Sc. degree in telecommunication engineering and the Ph.D. degree in signal processing from the Universitat Politecnica de Catalunya (UPC), Catalunya, Spain, in 1986 and 1990, respectively. From 1986 to 1988, he was involved in his doctoral work at the Department of Signal Theory and Communications, UPC.

From 1989 to 1990, he worked as an Assistant Professor and from 1991 to 1993, he worked as an Associated Professor in the Department of Signal Theory and Communications, UPC. From February 1995 to January 1996, he was with AT&T Bell Laboratories, Murray Hill, NJ, as a Consultant in speech recognition. Currently, he is a Full Professor of signal theory and communications in the Department of Electronic Engineering and Communications, University of Zaragoza, Zaragoza, Spain, where he is heading a research team in speech recognition and signal processing. He has managed several speech-related projects in Spain. He has coauthored more man 100 technical papers in the field of speech and speaker recognition, speech enhancement and recognition in adverse acoustic environments, acoustic modeling, confidence measures, and spoken dialogue systems.

Richard Rose (SM'00) received the B.S. and M.S. degrees in electrical and computer engineering from the University of Illinois, Urbana-Champaign, and the Ph.D. degree in electrical engineering from the Georgia Institute of Technology, Atlanta.

From 1980 to 1984, he was with Bell Laboratories working on signal processing and digital switching systems. From 1988 to 1992, he was with MIT Lincoln Laboratory working on speech recognition and speaker recognition. He was with AT&T Bell Laboratories from 1992 to 1996, and with AT&T Labs—Research, Florham Park, NJ, from 1996 to 2003. Currently, he is an Associate Professor of electrical and computer engineering at McGill University, Montreal, QC, Canada.

Prof. Rose served as a member of the IEEE Signal Processing Society (SPS) Technical Committee on Digital Signal Processing from 1990 to 1995. He was elected as an at-large member of the Board of Governors for the SPS during the period from 1995 to 1997. He served as an Associate Editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING from 1997 to 1999. He served as a member of the IEEE SPS Speech Technical Committee (STC) from 2002 through 2005, and was the founding editor of the STC Newsletter. He also served as one of the general chairs of the 2005 IEEE Automatic Speech Recognition and Understanding Workshop. He is a member of Tau Beta Pi, Eta Kappa Nu, and Phi Kappa Phi.

Luis Buera was bom in Lleida, Spain, in 1978. He received the M.Sc. degree in telecommunication engineering from the University of Zaragoza (UZ), Zaragoza, Spain, in 2002. He is currently working towards the Ph.D. degree at UZ. His research is based on feature vector normalization techniques for robust speech recognition.

From 2002 to 2007, he was with the Communications Technologies Group, Department of Electronic Engineering and Communications, UZ, under a research grant.

Óscar Saz was born in Zaragoza, Spain, in 1980. He received the M.Sc. degree in telecommunication engineering from the University of Zaragoza (UZ), Zaragoza, Spain, in 2004. He is currently working towards the Ph.D. degree at UZ.

Since 2004, he has been with the Communications Technologies Group, Department of Electronic Engineering and Communications, UZ, under a research grant. Currently, his research interests are in the field of speaker adaptation and personalization, with a special focus on users suffering from speech impairments.

Alfonso Ortega was born in Teruel, Spain, in 1976. He received the M Sc. degree in telecommunication engineering and the Ph.D. degree from the University of Zaragoza (UZ), Zaragoza, Spain, in 2000 and 2005, respectively.

In 1999 he joined, under a research grant, the Communications Technologies Group, Department of Electronic Engineering and Communications, UZ, where he has been an Assistant Professor since 2001. He is also involved as a Researcher with the Aragon Institute of Engineering Research (13A), UZ. Currently, his research interest lies in the signal processing field applied to speech technologies.