Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language 21 (2007) 72–87
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Speaker-adaptive learning of resonance targets in a hidden trajectory model of speech coarticulation

Dong Yu *, Li Deng, Alex Acero

Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA

Received 2 November 2004; received in revised form 17 August 2005; accepted 22 December 2005
Available online 3 February 2006

Abstract

A novel speaker-adaptive learning algorithm is developed and evaluated for a hidden trajectory model of speech coarticulation and reduction. Central to this model is the process of bi-directional (forward and backward) filtering of the vocal tract resonance (VTR) target sequence. The VTR targets are key parameters of the model that control the hidden VTR's dynamic behavior and the subsequent acoustic properties (those of the cepstral vector sequence). We describe two techniques for training these target parameters: (1) speaker-independent training that averages out the target variability over all speakers in the training set; and (2) speaker-adaptive training that takes into account the variability in the target values among individual speakers. The adaptive learning is applied also to adjust each unknown test speaker's target values towards their true values. All the learning algorithms make use of the results of accurate VTR tracking as developed in our earlier work. In this paper, we present details of the learning algorithms and the analysis results comparing speaker-independent and speaker-adaptive learning. We also describe TIMIT phone recognition experiments and results, demonstrating consistent superiority of speaker adaptive learning over speaker-independent one measured by the phonetic recognition performance.

1. Introduction

In recent years, much research in spoken language technology has been devoted to incorporating structures of human speech and language into statistical speech recognition systems, and a growing body of literature on this theme is emerging (e.g., Bakis, 1991; Chelba and Jelinek, 2000; Bilmes, 2004; Deng and Braam, 1994; Bridle et al., 1998; Deng, 1998; Deng et al., 2004a; Holmes and Russell, 1999; Rose et al., 1996; Ostendorf et al., 1996; Sun and Deng, 2002; Gao et al., 2000; Wang et al., 2004; Wegmann et al., 1996; Zhou et al., 2003). Researchers have explored the approaches of using the hidden structure of speech in the human speech generation process, either implicitly or explicitly (Bakis, 1991; Bridle et al., 1998; Deng et al., 2004b; Deng, 1998,

* Corresponding author. Present address: Speech Research Group, Microsoft Research, 12429 107th PL NE, Kirkland, WA 98034, USA. Tel.: +1 425 707 9282; fax: +1 425 823 8659.
E-mail addresses: dongyu@microsoft.com (D. Yu), deng@microsoft.com (L. Deng), alexac@microsoft.com (A. Acero).

0885-2308/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2005.12.002

2004; Ma and Deng, 2003). One key component of these hidden dynamic modeling approaches is a target-filtering operation in some non-observable (i.e., hidden) domain (Deng, 1998, 2004; Zhou et al., 2003).

Recently, we developed a bi-directional target filtering approach to modeling speech coarticulation and context assimilated reduction (Deng et al., 2004b, 2006b). This hidden trajectory model functionally achieves both anticipatory and regressive coarticulation, while leaving the phonological units as the linear phonemic sequence and bypassing the use of more elaborate nonlinear phonological constructs as described in Deng (1998). The work reported in this paper extends the work in Deng et al. (2004b, 2006b) by providing an adaptive mechanism for learning the essential target parameters. Compared with the models in Bridle et al. (1998), Deng (1998) which proposed the use of infinite-impulse-response (IIR) filters for characterizing the hidden speech dynamics, our current model gives a significantly simpler finite-impulse-response (FIR) filter implementation in the specific domain of hidden vocal tract resonances (VTRs). The hidden resonances are mapped to observable cepstra as the acoustic parameters using a parameter-free analytical function [in contrast to neural networks as proposed in Bridle et al. (1998) and Deng (1998)], offering clear advantages in model implementation and in constructing automatic recognition systems that incorporate the speech structure. Both computation saving and model parameter saving are significant due to the elimination of neural networks and their learning.

One key set of parameters in the hidden trajectory model is the VTR targets, which are specific to each phone but are context independent. How to determine the values of these parameters is critical to the success of applying the model to speech recognition. The simplest way is to train a single set of VTR targets for all the speakers; i.e., in a speaker-independent manner. In this case, the training averages out the VTR targets' variability over all speakers in the training set. However, VTRs and their targets are related to the vocal tract length of the speaker, and hence they vary among speakers. A single set of VTR targets can produce the VTR trajectories that typically match well with data for some speakers, but not for other speakers. In the work reported in this paper, we have successfully developed a novel speaker-adaptive training algorithm that takes into account the VTR target variability among speakers. In essence, the algorithm makes use of the results of a high-accuracy VTR tracking technique that provides the information about the relative vocal tract lengths between a generic speaker (averaged over all speakers in the training set) and a specific speaker in either the training or test data. This philosophy is similar to the vocal tract length normalization (VTLN) techniques developed in the past for normalizing acoustic variabilities among speakers (Eide and Gish, 1996; Lee and Rose, 1998; Zhan and Westphal, 1997). The novelty of our algorithm is to apply such relative vocal tract length estimates directly to achieve accurate estimates of the VTR targets in hidden trajectory modeling.

The organization of this paper is as follows. In Section 2, we provide an overview of the hidden trajectory model formulated in the bi-directional FIR-based target filtering framework. In Section 3, the basic, speaker-independent training technique for VTR target parameter estimation is derived and described. The more effective, speaker-adaptive learning algorithm that adjusts the VTR target parameters for each individual speaker is presented in Section 4, where the issues of how to obtain the normalization factor and how to use the normalization factor are addressed. In Section 5, analysis and phone recognition experiments and results based on TIMIT database are shown, providing evidence for the effectiveness of the speaker-adaptive learning technique.

2. Hidden trajectory model: an overview

The hidden trajectory model presented in this paper consists of two stages. In Stage I, the model converts the VTR target sequence to the VTR trajectory by using the phone sequence hypothesis and the boundaries. In Stage II, the model converts the VTR trajectory into the cepstral trajectory with sub-phone dependent bias parameters. We now describe these two stages in more detail.

Stage I of the model is a bi-directional filtering process on the VTR target sequence, where each phone is associated with a unique target vector and timing. This gives rise to both forward and backward coarticulation, since it makes the VTR value at each time dependent on not only the current phone's VTR target but also on the VTR targets of the adjacent phones. This filtering process has been shown to give quantitative prediction of the magnitude of contextually assimilated reduction and coarticulation (Deng et al., 2004b).

The filtering operation is implemented by a slowly time-varying, FIR filter characterized by the following non-causal, vector-valued, impulse response function:

$$h_s(k) = \begin{cases} c\gamma_{s(k)}^{-k}, & -D < k < 0, \\ c, & k = 0, \\ c\gamma_{s(k)}^{k}, & 0 < k < D, \end{cases} \tag{1}$$

where k represents time frame, typically with a length of 10 ms each, $\gamma_{s(k)}$ is the *stiffness* parameter vector, one component for each VTR order. Each component of the stiffness vector is a positive real value between zero and one. The subscript $s(k)$ in $\gamma_{s(k)}$ indicates that the stiffness vector is dependent on the segment state $s(k)$ which varies over time. D in (1) is the uni-directional length of the impulse response, representing the temporal extent of coarticulation in the temporal direction, assumed for simplicity to be equal in length for the forward direction (anticipatory coarticulation) and the backward direction (regressive coarticulation).

In (1), c is the normalization constant to ensure that the filter weights add up to one. This is essential for the model to produce target undershooting, instead of overshooting. To determine c, we require that the filter coefficients sum to one:

$$\sum_{k=-D}^{D} h_s(k) = c \sum_{k=-D}^{D} \gamma_{s(k)}^{|k|} = 1. \tag{2}$$

For simplicity, we make the assumption that over the temporal span $-D \leqslant k \leqslant D$, the stiffness parameter's value stays approximately constant

$$\gamma_{s(k)} \approx \gamma_{s(0)}.$$

That is, the adjacent segments within the temporal span $2D + 1$ in length which contribute to the coarticulated home segment have the same stiffness parameter value as that of the home segment. Under this assumption, we simplify (2) to

$$c \sum_{k=-D}^{D} \gamma_{s(k)}^{|k|} \approx c\left[1 + 2\left(\gamma_{s(0)} + \gamma_{s(0)}^2 + \ldots + \gamma_{s(0)}^D\right)\right] = c\frac{1 + \gamma_{s(0)} - 2\gamma_{s(0)}^{D+1}}{1 - \gamma_{s(0)}}.$$

Thus

$$c\left(\gamma_{s(0)}\right) \approx \frac{1 - \gamma_{s(0)}}{1 + \gamma_{s(0)} - 2\gamma_{s(0)}^{D+1}}. \tag{3}$$

The input to the above FIR filter is the target sequence, which is a function of discrete time and jumps at the segments' boundaries. Mathematically, the input is represented as a sequence of step-wise constant functions with variable durations and heights:

$$T(k) = \sum_{i=1}^{P} \left[u\left(k - k_{s_i}^{\mathrm{l}}\right) - u\left(k - k_{s_i}^{\mathrm{r}}\right)\right] T_{s_i}, \tag{4}$$

where $u(k)$ is the unit step function, k_s^{r}, $s = s_1, s_2, \ldots, s_P$ the right boundary sequence of the segments in the utterance, and k_s^{l}, $s = s_1, s_2, \ldots, s_P$ are the left boundary sequence. In general, $k_{s_{i+1}}^{\mathrm{l}} = k_{s_i}^{\mathrm{r}}$ for $1 \leqslant i < P$. The difference of the two gives the duration sequence. T_s, $s = s_1, s_2, \ldots, s_P$ are the target vectors for segment s.

In the work presented in this paper, we assume that both left and right boundaries (and hence the durations) of all the segments in an utterance are known. For the training set, the boundaries are provided in TIMIT database. For the test set, where the current model is used to predict the VTR frequency trajectories, the boundaries in the target sequence input to the filter come from a recognizer's forced alignment results, on which our experimental results described in this paper are based.

Given the filter's impulse response and the input to the filter, the filter's output as the model's prediction for the VTR trajectories is the convolution between these two signals. The result of the convolution within the boundaries of the home segment s is

$$\hat{g}_s(k) = h_{s(k)} \otimes T(k) = \sum_{\tau} c\left(\gamma_{s(\tau)}\right) T_{s(\tau)} \gamma_{s(\tau)}^{|k-\tau|}, \tag{5}$$

where the input target vector and the filter's stiffness vector may take not only values associated with the current home segment, but also those associated with the adjacent segments since the time τ in (5) can go beyond the home segment's boundaries.

A sequential concatenation of all outputs $\hat{g}_s(k)$, $s = s_1, s_2, \ldots, s_P$ constitutes the model prediction of VTR trajectories for the entire utterance:

$$\hat{g}(k) = \sum_{i=1}^{P} \left[u\left(k - k_{s_i}^{\mathrm{l}}\right) - u\left(k - k_{s_i}^{\mathrm{r}}\right)\right] \hat{g}_{s_i}(k). \tag{6}$$

Note that the convolution operation above carried out by the filter in the model guarantees continuity of the trajectories at each junction of two adjacent segments, contrasting the discontinuous jump in the input to the filter at the same junction. This continuity is applied to all classes of speech sounds including consonantal closure. This provides the mechanism for coarticulation and VTR target undershooting in the current hidden trajectory model. Examples of the step-wise constant function of target sequence $T(k)$ and its smoothed, filtered function $g(k)$ will be given for real speech data in Section 5 (Figs. 5 and 6).

The Stage II of the hidden trajectory model is responsible for converting the VTR vector $\hat{g}(k)$ at each time frame k into a corresponding vector of LPC cepstra $o(k)$. The mapping, as has been implemented, is in a memoryless fashion (i.e., no temporal smoothing), and is statistical rather than deterministic. To describe this mapping function, we decompose the VTR vector g into a set of K resonant frequencies f and bandwidths b. That is, let

$$g = \begin{pmatrix} f \\ b \end{pmatrix},$$

where

$$f = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_K \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_K \end{pmatrix}.$$

Then the statistical mapping from VTRs to cepstra is represented by

$$o(k) = \Psi(\hat{g}_s(k)) + \mu_s + v_s(k), \tag{7}$$

where v_s is a segment-dependent, zero-mean Gaussian random vector following the Gaussian distribution of $N(v; 0, \Sigma_s)$, and μ_s is a subsegment-dependent bias vector for the nonlinear predictive function $\Psi(g_s)$.

In (7), the output of the mapping function $\Psi(\hat{g})$ has the following parameter-free, analytical form for its nth vector component (i.e., nth order cepstrum):

$$o_n = \frac{2}{n} \sum_{k=1}^{K} \mathrm{e}^{-\pi n \frac{b_k}{f_{\mathrm{samp}}}} \cos\left(2\pi n \frac{f_k}{f_{\mathrm{samp}}}\right) \tag{8}$$

where the speech signal sampling frequency $f_{\mathrm{samp}} = 16{,}000$ Hz is used for the TIMIT data in our experiments. A detailed derivation of this analytical form can be found in Deng et al. (2006a).

3. Speaker-independent learning of VTR target parameters

In this section, we describe speaker-independent training of the VTR target vectors T_s, which is a function of the phone segment s but is context independent. Given the results of VTR tracking $\bar{z}(k)$, which we developed in the earlier work of Deng et al. (2006a), the training is aimed to maximize the likelihood of such tracked VTR "data". Assuming the tracked VTR data obey a Gaussian distribution, where the mean vector

is the VTR trajectory $\hat{g}(k)$ generated from the Stage I of the hidden trajectory model, and the covariance matrix is denoted by Q_s. (Note that $\hat{g}(k)$ contains the VTR target parameters T_s, which are to be optimized.) Then the objective function for the training becomes

$$\log P = -0.5 Q_s^{-1} \sum_{k=1}^{K} [\bar{z}(k) - \hat{g}_s(k)]^2,$$

whose gradient is

$$\frac{\partial \log P}{\partial T_s} = Q_s^{-1} \sum_{k=1}^{K} [\bar{z}(k) - \hat{g}_s(k)] \frac{\partial \hat{g}_s(k)}{\partial T_s}. \tag{9}$$

Using (5), we compute the gradient on the right-hand side of Eq. (9) as

$$\frac{\partial \hat{g}_s(k)}{\partial T_s} = \sum_{\tau=\max(k-D,bd_l)}^{\min(k+D,bd_r)} C_{\gamma_s} \gamma_s^{|k-\tau|}, \tag{10}$$

where bd_l, and bd_r are the left and right boundaries for the current phone segment s. Note that this gradient is not a function of the target parameters T_s for the optimization; i.e., it is a constant with respect to T_s. Choosing this constant to be the inverse of the total frames of all tokens of segment s in the training data, we have the following gradient descent estimate for T_s:

$$\hat{T}_s^{n+1} = \hat{T}_s^n + \alpha \frac{\sum_{\text{tok}} \sum_{k=1}^{K_s^{\text{tok}}} \left\{ \bar{z}^{\text{tok}}(k) - \hat{g}_s^{\text{tok}}(k|\hat{T}_s^n) \right\}}{\sum_{\text{tok}} K_s^{\text{tok}}}, \tag{11}$$

where the trajectory function $\hat{g}_s^{\text{tok}}(k)$ is determined by the FIR filter's output, and K_s^{tok} is the duration of token 'tok' of segment s. Superscript n above denotes iteration number. In our implementation, the learning rate α in Eq. (11) is set to be one.

Note that in Eq. (11), the "data" $\bar{z}^{\text{tok}}(k)$ is computed from an existing VTR tracker, and $\hat{g}_s^{\text{tok}}(k|\hat{T}_s^n)$ is computed using the target $\hat{T}_s^n$ from the previous iteration n. Initialization of the target parameters $\hat{T}_s^0$ is based on modified target values of Klatt synthesizer (Klatt, 1980). The full set of these modified values are tabulated in Deng and O'Shaughnessy (2003). Note that the estimate of $\hat{T}_s$, upon convergence of the iterations, is assumed to be the same for all speakers.

4. Speaker-adaptive learning of VTR target parameters

4.1. Introduction

In previous sections, we presented the hidden trajectory model where the unobserved VTR trajectory is predicted (in model Stage I) with only the sequence of phones and their boundaries by filtering the VTR targets using the bi-directional FIR filters. We assumed that the targets are the same for all the speakers. However, due to the difference in the vocal tract length as well as in the geometry for different speakers (e.g., Naito et al., 2002; Zhan and Waibel, 1997; McDonough et al., 1998), the VTR targets for different speakers would differ. To incorporate this speaker-specific target parameters in the hidden trajectory model, we have developed an iterative speaker-adaptive training technique to estimate the generic and adapted speaker's VTR target parameters $\hat{T}_s^{\text{generic}}$. During the training, the difference in each speaker's VTR targets is taken into account by scaling the generic VTR targets $\hat{T}_s^{\text{generic}}$ (computed from the immediately previous iteration):

$$\hat{T}_{s,\text{spk}} = \beta^{\text{spk}} \cdot \hat{T}_s^{\text{generic}}, \tag{12}$$

where β^{spk} is the speaker-dependent normalization or scaling factor inversely proportional to the vocal tract length of speaker 'spk'. The dot operation above denotes an element-by-element multiplication. In our current implementation, β^{spk} is a vector, with each component for each corresponding VTR frequency component.

(Bandwidth components of the VTR are not scaled.) The same scaling is applied to each of separate test speakers in the recognition task to be described in Section 5. This strategy of adapting the target parameters to individual speakers is motivated by the popular technique of vocal tract length normalization for the acoustic data as published in Eide and Gish (1996), Kamm et al. (1995), McDonough et al. (1998), Pye and Woodland (1997), Wegmann et al. (1996), Welling et al. (1998), Zhan and Westphal (1997), Zhan and Waibel (1997), and Naito et al. (2002).

4.2. Computation of the normalization factor

We now describe how the normalization factor β^{spk} is computed in our adaptive training algorithm. Again, we make use of the results of our existing VTR tracker (Deng et al., 2006a), and follow the same assumption as in the work of vocal tract length normalization that the ratio of average formant values of two speakers is a good estimate of the two speakers' vocal tract lengths. Thus, the normalization factor can be simply computed by

$$\beta^{\text{spk}} = \frac{\langle \bar{z}^{\text{spk}} \rangle}{\langle \bar{z}^{\text{train}} \rangle}, \tag{13}$$

where $\langle \bar{z}^{\text{spk}} \rangle$ denotes the sample average of VTR frequencies over all frames in the utterance from a specific speaker 'spk' (the specific speaker either in the training or in the test data), and $\langle \bar{z}^{\text{train}} \rangle$ is the sample average of VTR frequencies over all frames in the full training set corresponding to the generic speaker. During our experiments, we found that the estimates of VTR frequencies of many consonants are less reliable than those of vowels. We thus limit our VTR frequency estimates to only those frames corresponding to the vowels in computing the normalization vector:

$$\beta^{\text{spk}} = \frac{\langle \bar{z}_v^{\text{spk}} \rangle}{\langle \bar{z}_v^{\text{train}} \rangle}, \tag{14}$$

where for the training data of TIMIT (as in our experiments), the vowel regions are labeled in the database, and for the test data, we use the vowel regions hypothesized for each item in the N-best list.

In our experiments, we discovered further that the wide range of vowel VTR or formant values creates undesirable biases in the estimate of the normalization factor in Eq. (14). To illustrate this problem, the average VTR frequency for an utterance that contains vowel tokens of only /aa/ will be significantly different than that with vowel tokens of only /iy/, even though both utterances are generated by the same speaker. This causes the estimates of the normalization factor according to (14) to be vastly different, even if they should ideally be the same for the same speaker. To solve this problem, we further refine the estimate of the normalization factor in (14) to the one using normalized vowel VTR frequencies $\langle \bar{z}_v \rangle$ averaged over all tokens:

$$\beta^{\text{spk}} = \frac{\langle \bar{z}_v^{\text{spk}} \rangle_n}{\langle \bar{z}_v^{\text{train}} \rangle_n}, \tag{15}$$

The numerator and denominator in Eq. (15) are the normalized average VTR frequencies for individual speaker (denoted by 'spk') and for all speakers in the training set (denoted by 'train'), respectively. They are computed by

$$\langle \bar{z}_v \rangle_n = \frac{\sum_{i=1}^{V} N_{v_i} \frac{\langle \bar{z}_{v_i} \rangle}{\langle \bar{z}_{v_i}^{\text{train}} \rangle}}{\sum_{i=1}^{V} N_{v_i}}, \tag{16}$$

where V is the number of different vowels in all utterances in the training set, and N_{v_i} is the number of frames of vowel v_i. That is, the averages are carried out over all frames of all vowels in the training set. The division in Eq. (16) by $\langle \bar{z}_v^{\text{train}} \rangle$ accomplishes the normalization. After the normalization, the utterances from the same speaker that contain different vowels (such as only /aa/ and only /iy/ in separate utterances) will produce approximately the same $\langle \bar{z}_v^{\text{spk}} \rangle_n$ and hence the same normalization factor β^{spk}. Note that according to Eq. (16), $\langle \bar{z}_v^{\text{train}} \rangle_n = 1$. Therefore, Eq. (15) is simplified to

$$\beta^{\text{spk}} = \langle \bar{z}_v^{\text{spk}} \rangle_n = \frac{\sum_{i=1}^{V} N_{v_i} \frac{\langle \bar{z}_{v_i} \rangle}{\langle \bar{z}_{v_i}^{\text{train}} \rangle}}{\sum_{i=1}^{V} N_{v_i}}. \tag{17}$$

4.3. Algorithm summary and discussion

With Eq. (17), the normalization factor for each individual speaker in either training data or in test data are determined. For the test utterance, the use of Eq. (12) (after $\hat{T}_s^{\text{generic}}$ is obtained as outlined in the next paragraph) effectively adapts the target parameters in the hidden trajectory model to that test speaker. And these adapted target parameters are used for recognition.

$\hat{T}_s^{\text{generic}}$ is learned from the training data in an iterative, speaker-adaptive manner. Initial VTR target vectors, $\hat{T}_s^{\text{generic}}(0)$, are provided by the speaker-independent training as described in the preceding section. Then, the normalization factor for each speaker in the training set is computed by Eq. (17) and the same value is used in each iteration for updating the generic target parameters. The speaker-adapted VTR target parameters are then obtained according to Eq. (12). These adapted target parameters are in turn used to predict the speaker-specific VTR trajectory using Stage I of the hidden trajectory model, giving rise to the quantity of

$$\hat{g}_{s,\text{spk}}\left(k|\beta^{\text{spk}} \cdot \hat{T}_s^{\text{generic}}(n)\right).$$

This is then compared with the tracked VTR "data" and the difference is averaged over all frames, all tokens, and all speakers. The final iterative training formula is

$$\hat{T}_s^{\text{generic}}(n+1) = \hat{T}_s^{\text{generic}}(n) + \frac{\sum_{\text{spk}} \sum_{\text{tok}} \sum_{k=1}^{K_{s,\text{spk}}^{\text{tok}}} \left\{ \bar{z}_{s,\text{spk}}^{\text{tok}}(k) - \hat{g}_{s,\text{spk}}^{\text{tok}}(k|\beta^{\text{spk}} \cdot \hat{T}_s^{\text{generic}}(n)) \right\}}{\sum_{\text{spk}} \sum_{\text{tok}} K_{s,\text{spk}}^{\text{tok}}} \tag{18}$$

which can be shown to be a maximum likelihood estimate in a similar way to the derivation of the speaker-independent training of Eq. (11).

The intuition of Eq. (18) is clear. The second term is the adjustment of the VTR targets by the amount that is equal to the per-frame difference between model-predicted VTR trajectory and the tracked VTR "data". The amount of adjustment diminishes as the model-predicted trajectory closely matches the data. We found in experiments that the likelihood for the tracked VTR "data" is always monotonically increasing over the iterations of the training according to Eq. (18). In practice, we use four iterations of the algorithm of Eq. (18) for the 462 speakers in TIMIT training data. Beyond four iterations, the increase of the likelihood becomes much less than the earlier iterations.

After $\hat{T}_s^{\text{generic}}$ is learned from the training data, the process of adaptation/normalization is carried out on a sentence-by-sentence basis for each test sentence. First, Eq. (17) is used to determine the normalization factor. Then, Eq. (12) is used to determine the adapted target values for the sentence.

5. Experiments and results

In this section, we present the results of analysis and phone recognition experiments using the TIMIT database. These results provide evidence for the effectiveness of the speaker-adaptive learning technique just described. In particular, we compare the hidden trajectory models with the VTR target parameters trained using speaker-independent and speaker-adaptive algorithms, as detailed in Sections 3 and 4, respectively. We first show the distributional results of the normalization factors for all the 462 speakers in TIMIT's training set. We then use a typical speech utterance to demonstrate that with the adaptively learned target parameters, both the VTR trajectory prediction (as the output of model Stage I) and the cepstral trajectory prediction (as the output of model Stage II) match real speech data much better than using speaker-independent training. We finally show the phonetic recognition results with N-best rescoring, further demonstrating the superiority of the speaker-adaptive learning.

5.1. Distributional results on the estimate of the normalization factor

Here, we show the distributional results on the estimate of the normalization factor vector, β^{spk}, component by component (F1 to F4), using the estimation formula of Eq. (17) described in Section 4. The results are obtained from a total of 462 speakers in the TIMIT training set. These distributional results are plotted in Figs. 1–4, respectively, for the four components of β^{spk} in terms of histograms. The results are plotted separately for the male and female populations of the data.

According to Eq. (12), the normalization factor measures the ratio of the VTR targets of the specific speaker to the "generic" speaker. This ratio is approximately the ratio of the two speakers' vocal tract lengths. Since the generic speaker's targets are computed from the full pool of speakers including both males and females (according to Eq. (18)), its VTR target values are in between those of a typical female and a typical male speaker. The distributional results, shown consistently in Figs. 1–4, illustrate that the estimates of the

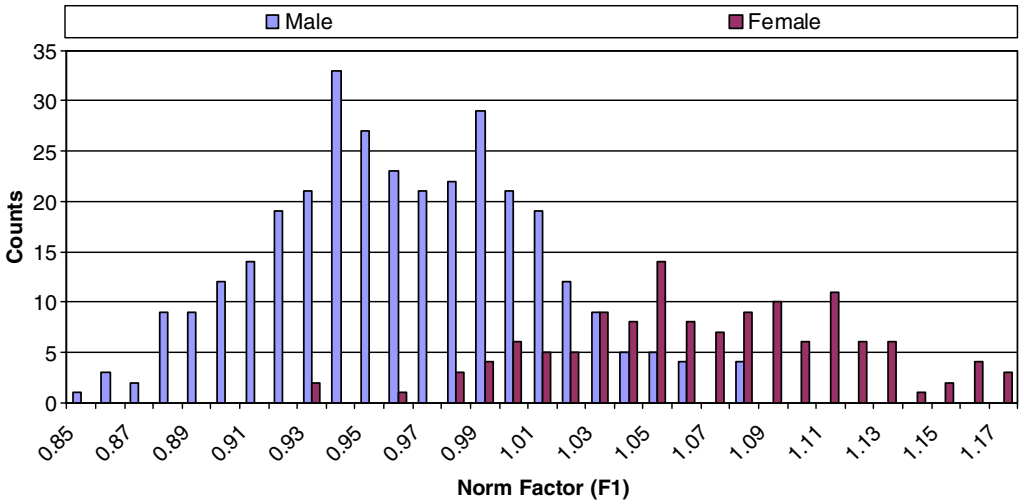

Fig. 1. Histograms of the estimate of the normalization factor for the F1 component. Data are from the 462 speakers in the TIMIT training set. Histograms for male and female speakers are plotted separately.

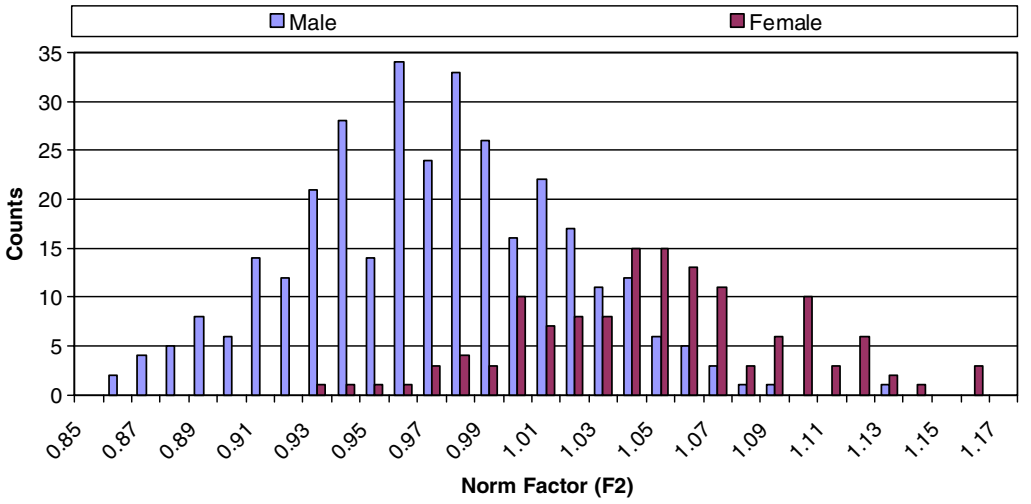

Fig. 2. Histograms of the estimate of the normalization factor for the F2 component. Data are from the 462 speakers in the TIMIT training set. Histograms for male and female speakers are plotted separately.

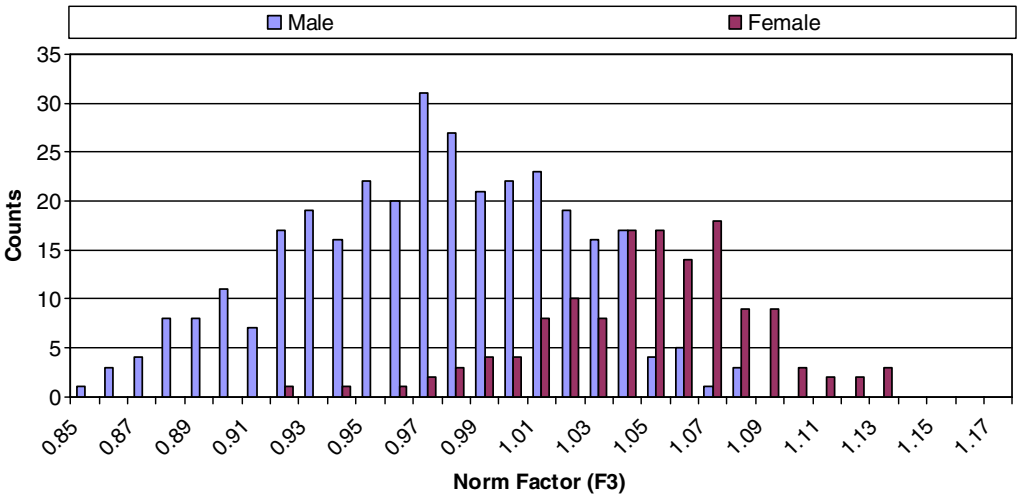

Fig. 3. Histograms of the estimate of the normalization factor for the F3 component. Data are from the 462 speakers in the TIMTT training set. Histograms for male and female speakers are plotted separately.

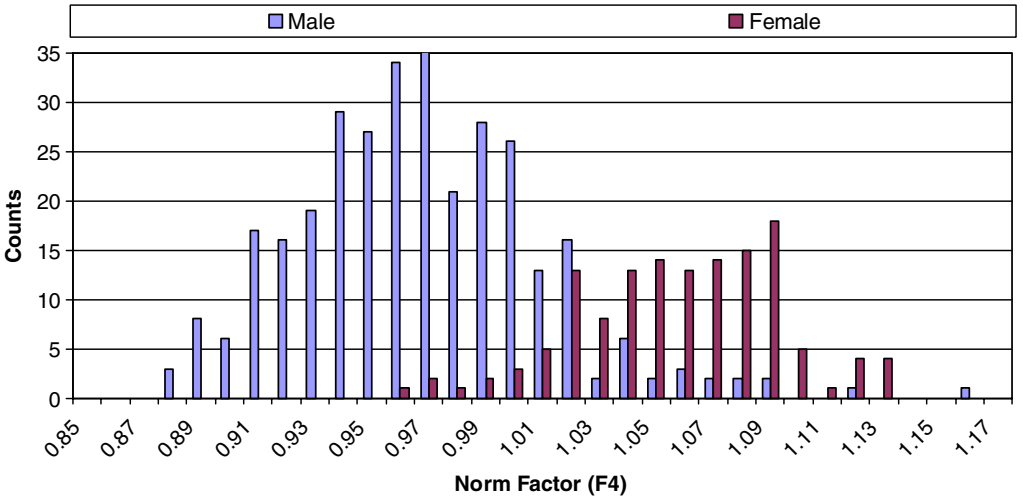

Fig. 4. Histograms of the estimate of the normalization factor for the F4 component. Data are from the 462 speakers in the TIMIT training set. Histograms for male and female speakers are plotted separately.

normalization factors for most of the female speakers are greater than one and those for most of the male speakers are lower than one. These are qualitatively consistent with the estimates of the relative vocal tract lengths using a completely different technique on the same TIMIT database (Zhan and Waibel, 1997). This consistency suggests that the estimation technique presented in Section 4 is effective.

The results in Figs. 1–4 illustrate that different speakers in the TIMIT database may have their VTR targets differ as much as 40% (ratio of two extreme values in the estimate of the normalization factor, 1.17/0.86). Male speakers tend to have higher normalization factors (or longer vocal tract lengths) while female speakers tend to have lower normalization factors (or shorter vocal tract lengths). If such significant differences are not taken into account, as in speaker-independent training, the VTR trajectory prediction based on the VTR targets (as inputs to the FIR filter) would be less accurate for those speakers that have the normalization factors substantially different from one. Inaccurate VTR trajectory prediction will lead to inaccurate cepstral sequence prediction, leading to greater error rates by the hidden trajectory model used as a speech recognizer. We will provide direct evidence for these points in the remainder of this section.

5.2. Results on prediction of VTR trajectories and cepstral sequences

We now demonstrate the effects of speaker-adaptive learning on the prediction accuracies of the VTR frequency trajectories (as the output of the hidden trajectory model Stage I) and of the cepstral sequences (as the output of the hidden trajectory model Stage II). In doing the predictions, the phone identities and their boundaries provided in the TIMIT database are used as the input to model Stage I. Since the hidden trajectory model assumes constant targets for each phone, we first decompose all compound phones (affricates and diphthongs) into their constituent sounds.

The model prediction incorporating VTR target adaptation proceeds as follows. The VTR targets for the generic speaker are trained using all training data according to Eq. (18), and for each individual speaker, we apply Eq. (17) to obtain the estimate of the normalization factor. The new adapted set of VTR targets for this speaker are then computed as the scaled version of the generic speaker's targets according to Eq. (12). These target values are then fed into the model Stage I for the VTR trajectory prediction. And the output of the predicted VTR trajectory is fed to model Stage II for the cepstral sequence prediction.

Fig. 5 shows the VTR prediction results for a female speaker, using the VTR targets obtained from the above speaker-adaptive learning. (Utterance SI487 by Speaker FKSRO in dialect region 7: "*Cable confirmation, it said translated*".) The prediction results, shown as four smooth lines (from F1 to F4), are superimposed on the spectrogram which shows true VTR trajectories (for the vocalic regions) as the spectral prominences or dark bands. These lines are the outputs of the FIR filter in model Stage I, and to illustrate the filter operation, the input to the filter is also plotted as the step-wise dashed lines representing the corresponding four VTR target sequences. For the majority of the frames, the filter's output either coincides or is close to the true VTR frequencies.

As a contrast, in Fig. 6, we show the same kind of VTR prediction results for the same speaker and utterance, but using the VTR targets obtained from speaker-independent training. The predicted VTR frequency values are now almost always lower than the true values as identified from the vocalic regions of the spectrogram. This comparison demonstrates the effectiveness of the target adaptation.

Further evidence for the effectiveness of speaker-adaptive target estimation is provided from the comparative results of cepstral sequence predictions as the output of model Stage II; i.e., using the nonlinear mapping from VTRs to cepstra in Eq. (8) on a frame-by-frame basis. Figs. 7 and 8 show such prediction (solid lines for C1, C2, and C3) with and without the speaker-adaptive learning, respectively. The inputs to the mapping function are the predicted VTR trajectories (by model Stage I) in Figs. 5 and 6, respectively. The dotted lines in

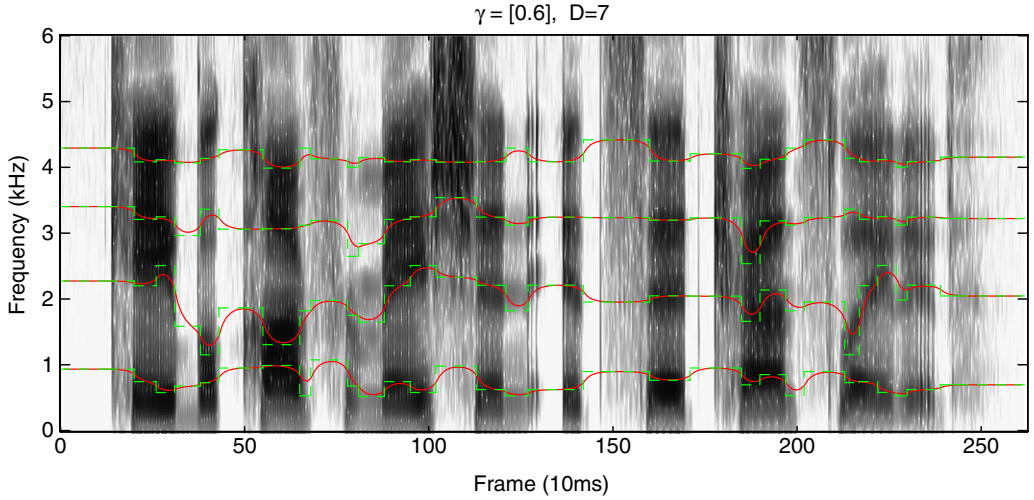

Fig. 5. VTR trajectory prediction using the bi-directional FIR filter and the VTR targets estimated with speaker-adaptive learning. The step-wise dashed lines are the target sequences (F1–4) as inputs to the FIR filter, and the continuous lines are the outputs (F1–4) of the filter as the predicted VTR frequency trajectories.

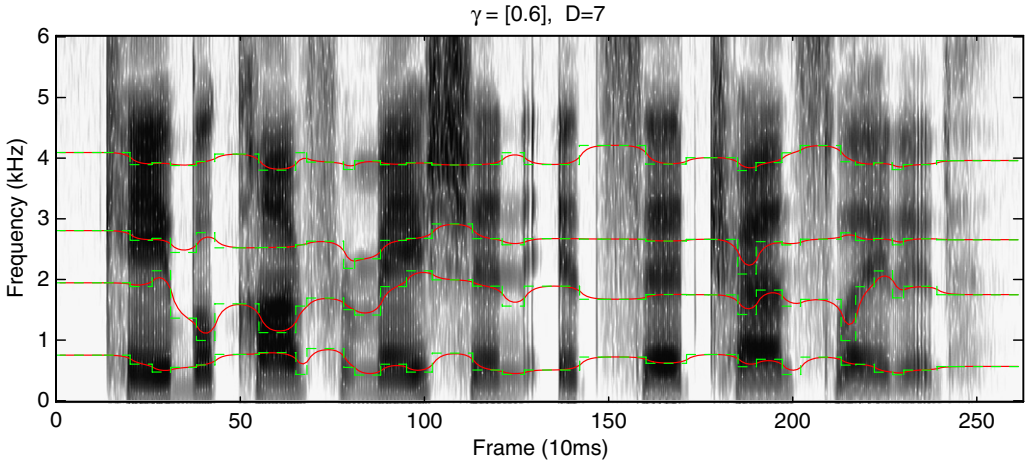

Fig. 6. VTR trajectory prediction using the bi-directional FIR filter, where the VTR targets are obtained from speaker-independent training. The step-wise dashed lines are the target sequences (F1–4) as inputs to the FIR filter, and the continuous lines are the outputs (F1–4) of the filter as the predicted VTR frequency trajectories.

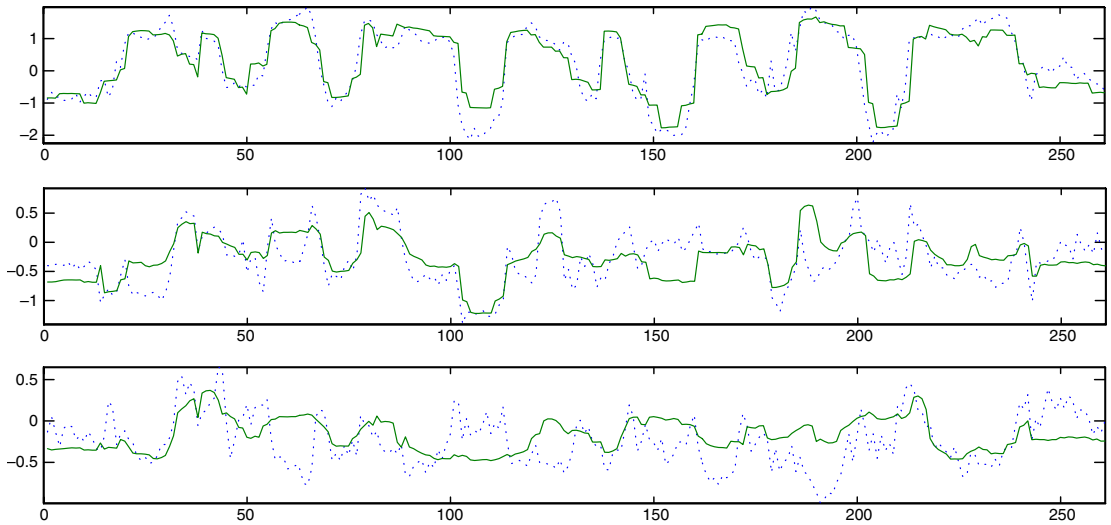

Fig. 7. Cepstral sequence prediction (solid lines) as the output of model Stage II (nonlinear mapping from VTRs to cepstra). The input to the mapping function is the VTR trajectory prediction of Fig. 5 with speaker-adapted VTR targets. The dotted lines are the LPC cepstra C1–C3 calculated from the acoustic signal waveform directly.

Figs. 7 and 8 are the LPC cepstral data C1–C3 calculated from the acoustic signal waveform directly. It is clear that the predicted cepstra in Fig. 7 with target adaptation fit the data more closely than those in Fig. 8. Since the model-to-data match at the acoustic level, calculated as the likelihood of the model evaluated on the observed acoustic data, is the criterion by which the speech recognition decision is made, the above comparative results on cepstral fitting between the speaker-adaptive and speaker-independent training suggest that better recognition performance can be achieved by the former than the latter. This is confirmed by the phonetic recognition experiment presented in Section 5.

The above results gave qualitative evidence that speaker adaptation improves both the VTR estimates and LPC cepstral prediction. We now present quantitative results based on a sample size of 192 TIMIT utterances in the core test set. We desire to show how much better, quantified in terms of the root mean square (RMS) error, speaker adaptation reduces the errors to the true VTR values averaged over these utterances. However,

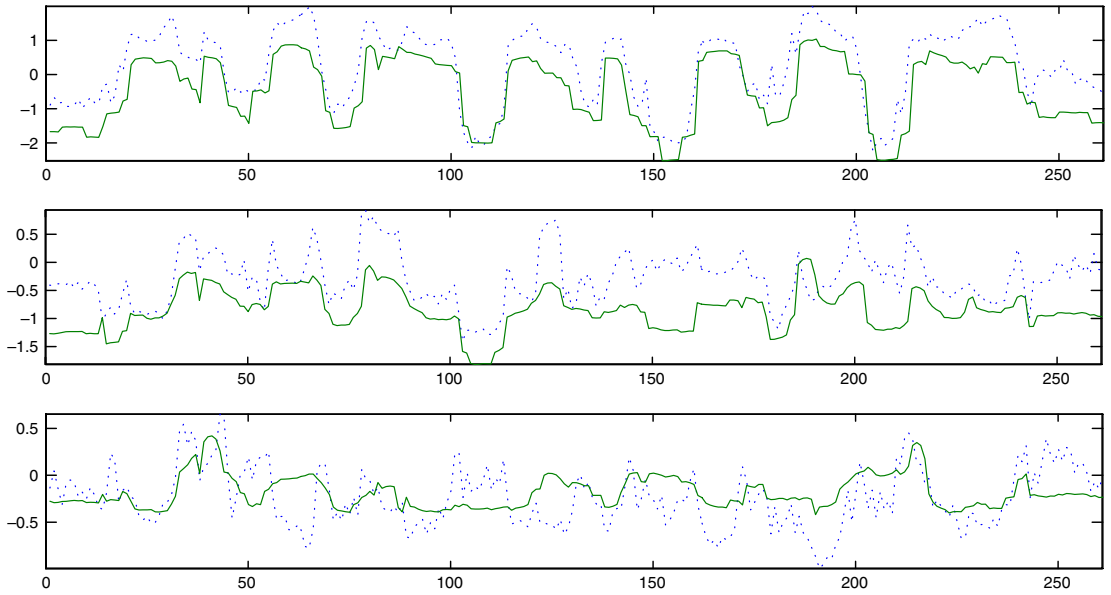

Fig. 8. Cepstral sequence prediction (solid lines) as the output of model Stage II (nonlinear mapping from VTRs to cepstra). The input to the mapping function is the VTR trajectory prediction of Fig. 6 with speaker-independently trained VTR targets. The dotted lines are the LPC cepstra C1–C3 calculated from the acoustic signal waveform directly.

since there is no such a reference database available, we use the results of the high-quality VTR tracking algorithm (Deng et al., 2004a) to approximate the reference VTR values for computing the RMS errors. (This is the same tracking algorithm used in the training phase.) In Table 1, we show the comparative results of the averaged RMS errors before and after speaker adaptation of the resonance targets. Four components of the VTR frequencies are listed separately. We observe a dramatic decrease of the RMS errors, especially for high-order VTR components. The greater errors when no adaptation is performed are due to the use of generic, speaker-independent targets as the input for the filtering. The error reduction after target adaptation to the individual speakers demonstrates the effectiveness of the adaptation algorithm we have developed.

Similar comparative results (also averaged over the 192 utterances) for LPC cepstra are obtained and shown in Table 2. Here, the reference LPC cepstra are computed directly from the acoustic waveforms, and are not approximated. Again, we observe the reduction of errors, mainly for low-order cepstra, due to the adaptation.

Table 1
Comparison of root mean square errors for VTR trajectories before and after speaker adaptation of the resonance targets

	F1 (Hz)	F2 (Hz)	F3 (Hz)	F4 (Hz)
RMS error with no adaptation	49	569	1024	1626
RMS error after adaptation	41	108	115	114

VTR references are approximated by a high-quality automatic VTR tracking algorithm. Results are averaged over 192 TIMIT utterances.

Table 2
Comparison of root mean square errors for LPC cepstral prediction before and after speaker adaptation of the resonance targets

	C2	C4	C6	C8	C10	C12
RMS error with no adaptation	0.51	0.18	0.14	0.12	0.08	0.07
RMS error after adaptation	0.20	0.15	0.13	0.10	0.08	0.07

Cepstral references are computed from the speech waveforms. Results are averaged over 192 TIMIT utterances.

5.3. Results on phonetic recognition

The experimental results of phonetic recognition on TIMIT with the standard core test set (192 utterances) are presented in this section to compare the relative effectiveness of the speaker-adaptive and speaker-independent training. In all the experiments reported in this section, we used the VTR tracker developed and reported in Deng et al. (2004a). Three states per segment are used for cepstral residual modeling. The coarticulation temporal extent parameter is set at $D = 7$. Variance parameters are estimated from sample variances from the data. The results are scored based on the standard mapping of the TIMIT phone set.

Due to the high computational cost of direct decoding using the long-span, wide-context hidden trajectory model, we limit our experiments to the N-best rescoring paradigm. For each of the N-best lists consisting of the hypothesized phone sequence and the constituent phone boundaries, the VTR trajectory $\hat{g}(k)$ is generated using the targets learned in the training or adaptation phase presented earlier in this paper. Then, the likelihood of the acoustics in terms of cepstral sequence corresponding to the hypothesized phone sequence and to the generated VTR trajectory is computed using Eq. (7).

We use a large-scale N-best list in the rescoring experiments in order to obtain meaningful results. In our experiments, the N-best list with $N = 1000$ is used, which is generated by a conventional, high-quality triphone HMM with phone bigram as the "language model". We found that even with the size of N to be as large as 1000, the oracle phone error rate (PER) is still over 18%. And increasing N to 2000 only reduces the oracle PER to 17%, while substantially increasing the computational cost of our N-best rescoring experiments. Such a high oracle error rate is not favorable to our long-span contextual hidden trajectory model, since any local error in the hypothesized phone(s) tends to propagate to its neighbors due to the continuity constraint across phones on the VTR trajectory represented in model Stage I. (This kind of "error propagation" effect is minimal for the short-span contextual models such as HMMs.) One simple way to artificially remove the error propagation effect is to manually add the reference hypothesis into the N-best list to form a new $N + 1 = 1001$ candidate list. A good model should be able to rank the reference hypothesis higher to the top among all the 1001 hypotheses than a poor model, reducing the "sentence" error rate (SER). This SER can serve a meaningful performance measure for the quality of a long-span wide-context model such as our hidden trajectory model.

We use this SER as the performance measure on the TIMIT core test set to compare the hidden trajectory model whose VTR targets are adapted to each speaker (Section 4) versus the same model but with the VTR targets trained speaker independently (Section 3). The results are shown in Fig. 9, where the oracle SER is plotted as a function of top number of choices in the overall 1001 hypotheses. The reference is always within the total 1001 candidates, no matter how they are re-scored. Hence, when the top number of candidates is increased to 1001, the (oracle) SER naturally becomes zero. However, when the top number of candidates varies below 1001, the hidden trajectory model with speaker adapted VTR targets is shown to consistently outperform the counterpart with no such adaptation. For example, within the top 50 candidates (among 1001 in total) after rescoring by the model with speaker adaptive learning, 40% of the utterances (among 192 in total) have the respective references included, whereas with the rescoring done by the model with speaker independent targets, the inclusion rate drops to 21% (i.e., 79% SER).

Similar results are obtained when the oracle PER, instead of oracle SER, is plotted in Fig. 10 for the two ways of determining the VTR targets. Again, consistently over all top numbers of candidates, the oracle PER is significantly lower for the speaker-adaptive model than for the speaker-independent model. Take the same data point at the top 50 candidates after rescoring by the model with speaker adaptive learning. Now the 40% of the reference inclusion rate (60% SER) gives 13% oracle PER. This is significantly lower than the 17% PER obtained by the model with the VTR targets trained speaker independently. For the top one candidate, the PER is dropped from 24% to 21% by speaker adaptive learning of targets, giving a 14% relative phone error reduction.

We emphasize that the above experiments were carried out by rescoring on the augmented $N = 1001$ best list with correct references included. The oracle errors for different depths of the single $N = 1001$ list are shown, which are generally not the top-best results. The main point of this paper is to compare the relative results with and without speaker adaptation, and not to achieve the best possible results; our follow-up work has been focused on this latter goal.

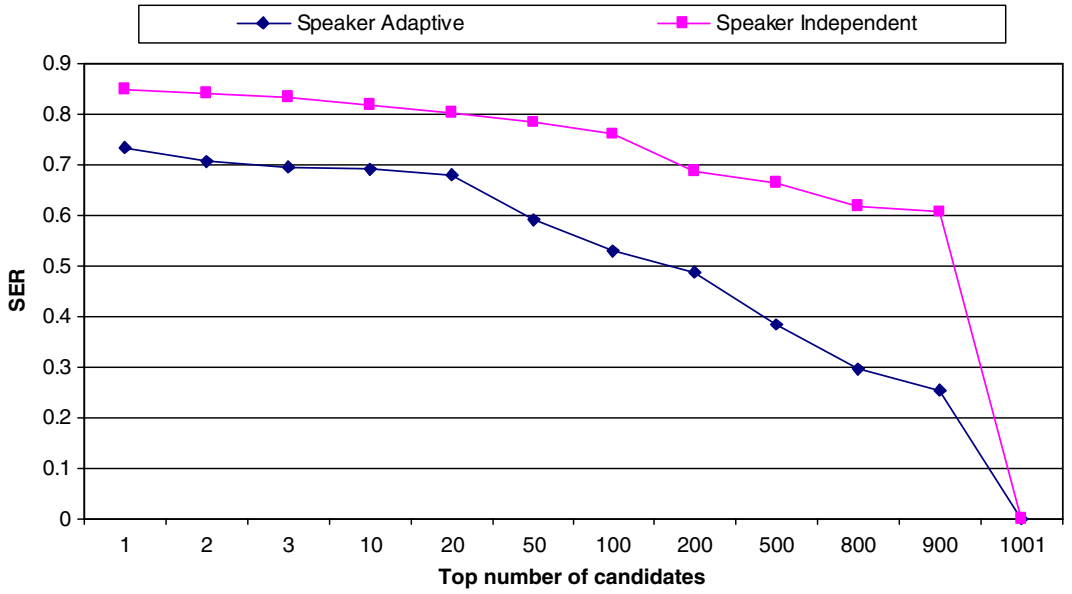

Fig. 9. Oracle sentence error rate (SER) on the TIMIT core test set (192 utterances) using the bi-directional target-filtering hidden trajectory model with its VTR targets determined in two ways: speaker adaptive learning vs. speaker independent learning. The result is from rescoring of 1000-best lists generated from a conventional tri-phone HMM, and with the correct hypothesis added. No language model is used in rescoring.

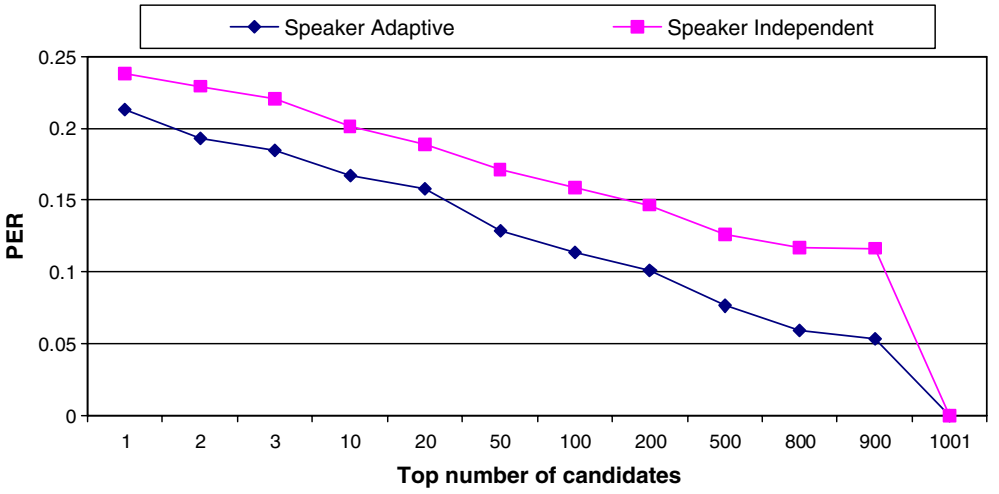

Fig. 10. Oracle phone error rate (PER) on the TIMIT core test set. Conditions are the same as in Fig. 9.

It is of interest to note that when un-augmented $N = 1000$ lists are used for the rescoring experiment, the effectiveness of the resonance target algorithm is reduced by about half. Shown in Table 3 are comparative PERs before and after applying the speaker adaptation algorithm (vertical), and also before and after aug-

Table 3
Phone error rate comparison: before and after speaker target adaptation, and with and without augmenting N-best lists by references

	Un-augmented N-best lists	Augmented N-best lists
PER with no adaptation	29.4%	24.1%
PER after adaptation	28.2%	21.0%

Results are obtained by rescoring using the N-best lists generated from an HMM system.

menting the *N*-best lists by adding references (horizontal). (No language model scores are used in the rescoring since we are concerned mainly with acoustic modeling.) The *N*-best lists are generated by a high-quality HMM system with a bi-phone language model, which gives the top-one PER of 27.5% (and SER of 100%). We note that while the speaker adaptation algorithm reduces the PER in the experiments involving un-augmented *N*-lists, the overall accuracy is low also. This is likely due to the "error propagation" effects that we discussed earlier in this subsection. Our future work is aimed to overcome such effects and the work presented in this paper has been limited to demonstrating the effectiveness of the speaker-adaptive algorithm.

6. Summary and conclusions

In this paper, we present a quantitative, two-stage model for predicting the VTR trajectories and then the subsequent cepstrum trajectories. This hidden trajectory model is based on a bi-directional filtering of phone-dependent, VTR target sequences implemented with a temporally symmetric FIR digital filter. The output of the filter is mapped to the LPC cepstrum via a parameter-free, analytical nonlinear prediction function. Given the LPC cepstral data computed from the input speech waveform, the likelihood that such acoustic data are generated from the model can be computed as the basis for speech recognition.

One most important set of parameters in the hidden trajectory model is the VTR targets, which drive the entire generative process of the VTR and acoustic trajectories. The main contribution of the work reported in this paper is the development of learning algorithms to automatically determine these target values from the observation data. Two algorithms are described in the paper. First, speaker-independent training is presented in Section 3, which is based on the simplifying assumption that a single set of VTR targets are associated with all the speakers. Second, speaker-adaptive learning is presented in Section 4, resulting in a speaker specific set of VTR target parameters.

Experiments are conducted and reported in this paper to demonstrate the critical role of the VTR target parameters in the model construction and its operation, and the superior performance of speaker adaptive learning. The distributional results of the normalization factors computed in the adaptive learning algorithm for all the 462 training speakers in TIMIT database are shown to demonstrate the wide variation of the VTR targets over speaker, both within and across genders. Typical speech utterances are then used to demonstrate that with the adaptively learned target parameters, both the VTR trajectory prediction and the cepstral trajectory prediction match real speech data much better than using speaker-independent training. The results show visually how the accuracy of the VTR and cepstral trajectory predictions depends upon the accuracy of the VTR targets. Further, a phonetic recognizer is constructed using the hidden trajectory model with two ways of determining the VTR target parameters. The recognizer is evaluated in a TIMIT phonetic recognition task and large-scale *N*-best rescoring paradigm is used for the evaluation. The results demonstrate a 14% phone error rate reduction using the model with speaker adaptation of targets compared with without such adaptation.

The future direction of the research in this area involves further expansion of the capability of the hidden trajectory model by developing parametric forms of the target distribution as related to speaker variation and speaking style variation. Our goal is to develop novel acoustic models of speech, exemplified by the hidden trajectory model presented in this paper, to jointly account for variations of speech due to context, speaker, speaking rate and style. This is to form a basis for building new types of conversational speech recognizers more effective than the conventional HMM technology.

Acknowledgements

We thank Dr. Mike Seltzer who built the triphone HMM system for generating a high-quality *N*-best list for rescoring the new model as reported in this paper.

References

Bakis, R., 1991. Coarticulation modeling with continuous-state HMMs. In: Proceedings of the IEEE Workshop Automatic Speech Recognition, Harriman, New York, pp. 20–21.

Bilmes, J., 2004. Graphical models and automatic speech recognition. In: Johnson, M., Ostendorf, M., Khudanpur, S., Rosenfeld, R. (Eds.), Mathematical Foundations of Speech and Language Processing. Springer, New York, pp. 135–186.

Bridle, J., Deng, L., Picone, J., et al., 1998. An investigation of segmental hidden dynamic models of speech coarticulation for automatic speech recognition. Final Report for the 1998 Workshop on Language Engineering, Center for Language and Speech Processing at Johns Hopkins University, pp. 1–61.

Chelba, C., Jelinek, F., 2000. Structured language modeling. Computer Speech Lang. (October), 283–332.

Deng, L., 1998. A dynamic, feature-based approach to the interface between phonology and phonetics for speech modeling and recognition. Speech Commun. 24 (4), 299–323.

Deng, L., 2004. Switching dynamic system models for speech articulation and acoustics. In: Johnson, M., Ostendorf, M., Khudanpur, S., Rosenfeld, R. (Eds.), Mathematical Foundations of Speech and Language Processing. Springer, New York, pp. 115–134.

Deng, L., Braam, D., 1994. Context-dependent Markov model structured by locus equations: applications to phonetic classification. J. Acoust. Soc. Am. 96, 2008–2025.

Deng, L., O'Shaughnessy, D., 2003. Speech Processing – A Dynamic and Optimization-Oriented Approach. Marcel Dekker, New York, NY.

Deng, L., Lee, L., Attias, H., Acero, A., 2004a. A structured speech model with continuous hidden dynamics and prediction-residual training for tracking vocal tract resonances. In: IEEE Proceedings of ICASSP, May 2004, vol. I, pp. 557–560.

Deng, L., Yu, D., Acero, A., 2004b. A quantitative model for formant dynamics and contextually assimilated reduction in fluent speech. ICSLP 2004, Jeju, Korea.

Deng, L., Acero, A., Bazzi, I., 2006a. Tracking vocal tract resonances using a quantized nonlinear function embedded in a temporal constraint. IEEE Trans. Speech Audio Process 14 (2), in press.

Deng, L., Yu, D., Acero, A., 2006b. A bi-directional target-filtering model of speech coarticulation and reduction: two-stage implementation for phonetic recognition. IEEE Trans. Speech Audio Process 14 (1), 256–265.

Eide, E., Gish, H., 1996. A parametric approach to vocal tract length normalization. In: IEEE Proceedings of ICASSP, pp. 346–348.

Gao, Y., Bakis, R., Huang, J., Zhang, B., 2000. Multistage coarticulation model combining articulatory, formant and cepstral features. In: Proceedings of ICSLP, vol. 1, pp. 25–28.

Holmes, W., Russell, M., 1999. Probabilistic-trajectory segmental HMMs. Computer Speech Lang. 13, 3–37.

Kamm, T., Andreou, G., Cohen, J., 1995. Vocal tract normalization in speech recognition: compensating for systematic speaker variability. In: Proceedings of the 15th Annual Speech Research Symposium. CLSP, Johns Hopkins University, Baltimore, MD, pp. 161–167.

Klatt, D., 1980. Software for a cascade/parallel formant synthesizer. J. Acoust. Soc. Am. 99 (3), 971–995.

Lee, L., Rose, R., 1998. A frequency warping approach to speaker normalization. IEEE Trans. Speech Audio Process. 6, 49–60.

Ma, J., Deng, L., 2003. Efficient decoding strategies for conversational speech recognition using a constrained nonlinear state-space model for vocal-tract-resonance dynamics. IEEE Trans. Speech Audio Process. 11, 590–602.

McDonough, J., Byrne, W., Luo, X., 1998. Speaker normalization with all-pass transforms. In: Proceedings of ICSLP, vol. 6, pp. 2307–2310.

Naito, M., Deng, L., Sagisaka, Y., 2002. Speaker clustering for speech recognition using vocal-tract parameters. Speech Commun. 36 (3–4), 305–315.

Ostendorf, M., Digalakis, V., Rohlicek, J., 1996. From HMMs to segment models: a unified view of stochastic modeling for speech recognition. IEEE Trans. Speech Audio Process. 4, 360–378.

Pye, D., Woodland, P.C., 1997. Experiments in speaker normalisation and adaptation for large vocabulary speech recognition. In: IEEE Proceedings of ICASSP, pp. 1047–1050.

Rose, R., Schroeter, J., Sondhi, M., 1996. The potential role of speech production models in automatic speech recognition. J. Acoust. Soc. Am. 99, 1699–1709.

Sun, J., Deng, L., 2002. An overlapping-feature based phonological model incorporating linguistic constraints: applications to speech recognition. J. Acoust. Soc. Am. 111 (2), 1086–1101.

Wang, W., Stolcke, A., Harper, M., 2004. The use of a linguistically motivated language model in conversational speech recognition. In: IEEE Proceedings of ICASSP, May 2004.

Wegmann, S., McAllaster, D., Orloff, J., Peskin, B., 1996. Speaker normalization on conversational telephone speech. In: IEEE Proceedings of ICASSP, pp. 339–341.

Welling, L., Haeb-Umbach, R., Aubert, X., Haberland, N., 1998. A study on speaker normalization using vocal tract normalization and speaker adaptive training. In: IEEE Proceedings of ICASSP, Seattle, WA, May 1998, vol. 2, pp. 797–800.

Zhan, P., Waibel, A., 1997. Vocal tract length normalization for large vocabulary continuous speech recognition. CMU-CS-97-148, Carnegie Mellon University, Pittsburgh, PA, May 1997.

Zhan, P., Westphal, M., 1997. Speaker normalization based on frequency warping. In: IEEE Proceedings of ICASSP, pp. 1039–1042.

Zhou, J., Seide, F., Deng. L., 2003. Coarticulation modeling by embedding a target-directed hidden trajectory model into HMM. In: IEEE Proceedings of ICASSP, April 2003, vol. I, pp. 744–747.