A Bidirectional Target-Filtering Model of Speech Coarticulation and Reduction: Two-Stage Implementation for Phonetic Recognition

Li Deng, Dong Yu, and Alex Acero

***Abstract*—A structured generative model of speech coarticulation and reduction is described with a novel two-stage implementation. At the first stage, the dynamics of formants or vocal tract resonances (VTRs) in fluent speech is generated using prior information of resonance targets in the phone sequence, in absence of acoustic data. Bidirectional temporal filtering with finite-impulse response (FIR) is applied to the segmental target sequence as the FIR filter's input, where forward filtering produces anticipatory coarticulation and backward filtering produces regressive coarticulation. The filtering process is shown also to result in realistic resonance-frequency undershooting or reduction for fast-rate and low-effort speech in a contextually assimilated manner. At the second stage, the dynamics of speech cepstra are predicted analytically based on the FIR-filtered and speaker-adapted VTR targets, and the prediction residuals are modeled by Gaussian random variables with trainable parameters. The combined system of these two stages, thus, generates correlated and causally related VTR and cepstral dynamics, where phonetic reduction is represented explicitly in the hidden resonance space and implicitly in the observed cepstral space. We present details of model simulation demonstrating quantitative effects of speaking rate and segment duration on the magnitude of reduction, agreeing closely with experimental measurement results in the acoustic-phonetic literature. This two-stage model is implemented and applied to the TIMIT phonetic recognition task. Using the N-best ($N = 2000$) rescoring paradigm, the new model, which contains only context-independent parameters, is shown to significantly reduce the phone error rate of a standard hidden Markov model (HMM) system under the same experimental conditions.**

***Index Terms*—Cepstral dynamics, contextual assimilation, filtering of targets, formant dynamics, long-span context dependence, phonetic recognition, phonetic reduction, resonances, TIMIT.**

I. Introduction

THE IMPORTANCE of incorporating structures of human speech and language into statistical models for technology applications has been well known, and active research in this direction has been pursued in recent years [2], [3], [4], [6], [10], [11], [19], [27], [31], [33]. For speech recognition applications, the dynamic structure of human speech has been exploited in several ways in the past. Earlier work directly represented speech dynamics in the observed acoustic domain [5], [16], [19], [24]. More recent work explored the hidden structure of speech associated with various levels in the human speech generation process, either implicitly or explicitly [1], [3], [6], [8], [14], [22]. Common among these hidden dynamic modeling approaches is a target-filtering operation in some nonobservable domain. One directional (left-to-right) target filtering has been used in [6], [8], and [33]. It functionally approximates the causal physical system in speech articulation that accounts for inertia-related perseverance coarticulation, while the anticipatory coarticulation is modeled at the separate phonological level via the mechanism of nonlinear "atomic unit" overlapping or target look-ahead [8], [29]. The research reported in this paper simplifies the previous approach by merging the two separate levels of coarticulation modeling into the same level of the hidden dynamics, with bidirectional instead of uni-directional target filtering. This functionally achieves both anticipatory and regressive coarticulation, while leaving the phonological units as the linear phonemic sequence and bypassing the use of more elaborated nonlinear phonological constructs. This bidirectional filtering approach was originally proposed in [3], using a recursive, infinite-impulse response (IIR) filter with a high computational complexity. The current work presents a significantly simpler finite-impulse response (FIR) filter implementation of the hidden dynamics in the specific domain of vocal tract resonances (VTRs) or formants. In conjunction with the second-stage mapping from the hidden resonances to observable cepstra using a free-parameter analytical function (instead of a neural network as in [3]), the new two-stage model presented in this paper offers significant advantages in model implementation and in constructing automatic recognition systems that incorporate the hidden dynamic structure of speech.

A central component (Stage I) of the two-stage model presented in this paper is one that parsimoniously parameterizes the VTR dynamics within the bidirectional FIR target filtering framework. This is a joint model for coarticulation and reduction, both mediated by the hidden or unobserved VTR dynamics. Dynamic patterns of VTRs in fluent speech, especially those which are correlated with spectral prominences or formants for vowel sounds, have been a subject of intensive research in phonetics and in speech synthesis for many years [15], [17], [18], [23], [25], [26], [30], [32]. The research has been focusing on the central issue that the same formant values taken from the

Manuscript received July 20, 2004; revised November 17, 2004. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Timothy J. Hazen.

A portion of this work was presented at the ICSLP Conference 2004, Jeju Islands, Korea.

The authors are with Microsoft Research, Redmond, WA 98052 USA (e-mail: deng@microsoft.com; dongyu@microsoft.com; alexac@microsoft.com).

Digital Object Identifier 10.1109/TSA.2005.854107

1558-7916/$20.00 © 2006 IEEE

middle portion of a speech sound from its dynamic pattern can correspond to different sound classes specified solely in static terms. This inherent "static" confusion of speech classes without dynamic aspects of speech sound specification is believed to be one significant factor impeding current HMM-based speech recognition for casual-style, conversational speech. The VTR model in this paper gives dynamic specification of speech sounds, where the observed dynamic pattern of speech becomes the result of an interaction among phonetic context, speaking rate/duration, and spectral rate of change as related to speaking style [23]. In particular, our model assumes that each speech sound is specified by a largely context-independent target (but speaker dependent) in the VTR space, together with the *stiffness* parameter specifying how VTR trajectories may be formed in any given phonetic and prosodic environment. In the implementation of the model, the stiffness parameter is used to control temporal filtering of the sequentially arranged, VTR targets, and is dependent on a range of prosodic factors, speaking style in particular. The result of the temporal filtering, in both forward and backward directions, gives rise to the phonetically realized dynamic VTR patterns. A direct consequence of this filtering operation is as follows: the shorter a segment is, the greater the difference becomes between the filter's input as the target VTR values and the output as the actual VTR values. (Note the filter output also depends on the stiffness parameter associated with speaking style, in addition to the dependency on the filter input.) Therefore, our model naturally simulates the target-undershooting, or reduction phenomenon [18], [23], [25]. Because the input to the filter is the phonetically composed, discontinuous target sequence, which is smoothed by the filter resulting in continuous, "reduced" trajectories, this filter-based model (Stage I) represents the reduction phenomenon in a contextually assimilated manner. That is, reduction and coarticulation are jointly represented in the filter model. This type of model construction has been motivated by the reduction mechanism suggested originally in [18] and [23].

The organization of this paper as follows. In Section II, we provide mathematical details of Stage I of the overall two-stage model. Stage II of the model is described in Section III. This stage takes the VTR dynamics, which are the output of Stage I, as its input and produces the corresponding linear predictive coding (LPC) cepstral vector as its output on a frame-by-frame basis. Sections IV and V present simulation results for Stage I and Stage II components of the model, respectively, and comparisons are made between model prediction and acoustic measurements in real speech data. In constructing a phonetic recognizer using this two-stage model, the overall model's output in the form of the cepstral vector sequence is used as the observation for the recognizer. Specific issues in the recognizer design are discussed in Section VI, where experimental results using the N-best rescoring paradigm for the TIMIT phonetic recognition task are presented also. The results demonstrate the superior performance of the new system over the conventional HMM system. Finally, in Section VII, we discuss our future direction of research toward the goal of recognizing conversational speech, where a continuously varying degree of phonetic reduction and "static" sound confusion is captured by the fundamental mechanism of target filtering as presented in this paper.

II. Model Stage I: From Resonance Target Sequence to Resonance Dynamics

Stage I of the coarticulation and reduction model presented in this section is responsible for converting a sequence of VTR targets with discrete jumps at the phone segments' boundaries into the a smooth dynamic pattern (i.e., trajectory) across all these boundaries. Forward as well as backward coarticulation occurs when the bidirectional filtering and smoothing process makes the VTR value at each time dependent on not only the VTR target at the current phone, but also the VTR targets from the adjacent phones. In the mean time, the filtering process automatically exhibits contextually assimilated reduction when the segment's duration is reasonably short, especially when the filter parameter, which we call *stiffness*, of the filter is close to one. Reduction is defined in this paper as VTR target undershooting, i.e., the physically realized VTR value being away from the VTR target. When reduction is controlled by the targets of contextual (left and right) segments, we say that the reduction is contextually assimilated.

The model described in this section gives quantitative prediction of the magnitude of contextually assimilated reduction. It is constructed using a slowly time-varying, FIR filter characterized by the following noncausal, vector-valued, impulse response function:

$$\mathbf{h}_s(k) = \begin{cases} \mathbf{c}\boldsymbol{\gamma}_{s(k)}^{-k} & -D < k < 0 \\ \mathbf{c} & k = 0 \\ \mathbf{c}\boldsymbol{\gamma}_{s(k)}^{k} & 0 < k < D \end{cases} \tag{1}$$

where k represents time frame, typically with a length of 10 ms each. $\boldsymbol{\gamma}_{s(k)}$ is the "stiffness" parameter vector, one component for each resonance order. Each component is positive and real-valued, ranging between zero and one. In this paper, $\boldsymbol{\gamma}$ is treated as a deterministic quantity for simplicity purposes. (In a more comprehensive version of the model, $\boldsymbol{\gamma}$ is a Gaussian random vector characterized by the mean vector and covariance matrix.) The subscript $s(k)$ in $\boldsymbol{\gamma}_{s(k)}$ indicates that the stiffness parameter is dependent on the segment state $s(k)$ which varies over time. The multiplication of two vectors in (1) is on the component-by-component basis. D in (1) is the unidirectional length of the impulse response. It represents the temporal extent of coarticulation in one temporal direction, assumed for simplicity to be equal in length for the forward direction (anticipatory coarticulation) and the backward direction (regressive coarticulation).

In (1), $\mathbf{c}$ 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 $\mathbf{c}$, we require that the filter coefficients sum to one

$$\sum_{k=-D}^{D} \mathbf{h}_s(k) = \mathbf{c} \sum_{k=-D}^{D} \boldsymbol{\gamma}_{s(k)}^{|k|} = 1. \tag{2}$$

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

$$\boldsymbol{\gamma}_{s(k)} \approx \boldsymbol{\gamma}.$$

That is, the adjacent segments within the temporal span of $2D+1$ 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

$$\mathbf{c}\sum_{k=-D}^{D}\boldsymbol{\gamma}_{s(k)}^{|k|}\approx\mathbf{c}[1+2(\boldsymbol{\gamma}+\boldsymbol{\gamma}^2+\cdots+\boldsymbol{\gamma}^D)]=\mathbf{c}\frac{1+\boldsymbol{\gamma}-2\boldsymbol{\gamma}^{D+1}}{1-\boldsymbol{\gamma}}.$$

Thus

$$\mathbf{c}(\boldsymbol{\gamma})\approx\frac{1-\boldsymbol{\gamma}}{1+\boldsymbol{\gamma}-2\boldsymbol{\gamma}^{D+1}}. \tag{3}$$

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

$$\mathbf{t}(k)=\sum_{i=1}^{P}[\mathbf{u}(k-k_{s_i}^l)-\mathbf{u}(k-k_{s_i}^r)]\mathbf{t}_{s_i} \tag{4}$$

where $\mathbf{u}(k)$ is the unit step function, k_s^r, $s=s_1,s_2,\ldots,s_P$ are the right boundary sequence of the segments (P in total) in the utterance, and k_s^l, $s=s_1,s_2,\ldots,s_P$ are the left boundary sequence. Note the constraint on these starting and end times: $k_{s+1}^l=k_s^r$. The difference of the two boundary sequences gives the duration sequence. $\mathbf{t}_s$, $s=s_1,s_2,\ldots,s_P$ are the target vectors for segment s. (In a more comprehensive version of the model, the target vector values are drawn from a statistical distribution, whose parameters are automatically learned in a manner similar to [7]).

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 (e.g., those provided in TIMIT database). However, in general cases where the current model is used to predict the VTR frequency trajectories as the FIR filter's output, the boundaries in the target sequence input to the filter are not given. They either come from a recognizer's forced alignment results, on which our experimental results described in this paper are based, or need be learned automatically using advanced algorithms in a similar spirit to that described in [22].

Given the filter's impulse response and the input to the filter as described previously, 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

$$\mathbf{g}_s(k)=\mathbf{h}_{s(k)}*\mathbf{t}(k)=\sum_{\tau=k-D}^{k+D}\mathbf{c}(\boldsymbol{\gamma}_{s(\tau)})\mathbf{t}_{s(\tau)}\boldsymbol{\gamma}_{s(\tau)}^{|k-\tau|} \tag{5}$$

where the input target vector's value and the filter's stiffness vector's value may take not only those associated with the current home segment, but also those associated with the adjacent segments. The latter case happens when the time τ in (6) goes beyond the home segment's boundaries; i.e., when the segment $s(\tau)$ occupied at time τ switches from the home segment to an adjacent one.

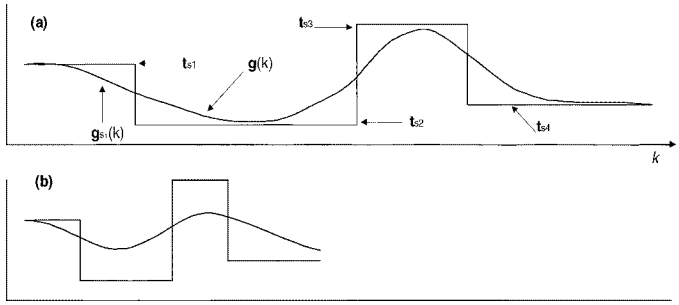

Fig. 1. Illustrations of the various VTR quantities in model Stage-I in an utterance with four phone segments. (a) and (b) are for the same four VTR targets and their filtered results, but the durations of the four segments are shorter in (b) than in (a).

A sequential concatenation of all outputs $\mathbf{g}_s(k)$, $s=s_1,s_2,\ldots,s_P$ in (5), each corresponding to a single segment in the utterance, constitutes the model prediction of VTR trajectories for the entire utterance

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

Note that the convolution operation 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.

The various VTR quantities in model Stage-I discussed previously are graphically illustrated in Fig. 1(a). Four segments, or $P=4$, are sequentially concatenated with their respective VTR targets, where one-dimensional VTR is used as the example for simplicity. The smoothed curve, $\mathbf{g}(k)$, is the result of FIR filtering, which runs over the entire duration of the four segments. The separate segment-bounded portions of the curve are denoted with subscript s. Fig. 1(b) shows the same VTR targets and their filtered results, but the durations of the segments are shorter.

III. Model Stage II: From Resonance Dynamics to Cepstrum Dynamics

We now present Stage II of the overall coarticulation and reduction model, which is responsible for converting the VTR vector $\mathbf{g}(k)$ at each time frame k into a corresponding vector of LPC cepstra $\mathbf{o}(k)$. Thus, the smooth dynamic pattern of $\mathbf{g}(k)$ as the output from Stage I is mapped to a dynamic pattern of $\mathbf{o}(k)$, which is typically less smooth, reflecting quantal properties in speech production [28]. 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 into a set of Q resonant frequencies $\mathbf{f}$ and bandwidth $\mathbf{b}$. That is, let

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

$$\text{where} \quad \mathbf{f} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_Q \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_Q \end{pmatrix}.$$

The choice of the highest resonance order Q in the current implementation of model Stage II is based on a compromise between the accuracy of the model prediction for data and the phonetically meaningful information contained in the resonances lower than order Q. The larger the Q is, the greater is the model prediction accuracy on speech acoustics but the less useful information is contained in the higher resonances pertaining to phonetic discrimination. This compromise leads to the empirical choice of $Q = 4$ in the current model implementation.

Then, the statistical mapping from VTRs to cepstra, which constitutes Stage II of the model, is represented by

$$\mathbf{o}(k) = \mathcal{O}(\mathbf{g}_s(k)) + \boldsymbol{\mu}_{ss} + \mathbf{v}_{ss}(k) \tag{7}$$

where $\mathbf{v}_{ss}$ is a subsegment-dependent, zero-mean Gaussian random vector: $\mathbf{v}_{ss} \sim \mathcal{N}(\mathbf{v}; \mathbf{0}, \boldsymbol{\Sigma}_{ss})$, and $\boldsymbol{\mu}_{ss}$ is a subsegment-dependent bias vector for the nonlinear predictive function $\mathcal{O}(\mathbf{g}_s)$. A subsegment of a phone is defined to be a consecutive temporal portion of the phone segment. Linear concatenation of several subsegments constitutes a phone segment.

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

$$o_n = \frac{2}{n} \sum_{q=1}^{Q} e^{-\pi n (b_q / f_s)} \cos\left(2\pi n \frac{f_q}{f_s}\right) \tag{8}$$

where f_s denotes sampling frequency of the speech signal. For TIMIT data which we have used in experiments, we have $f_s =$ 16000 Hz. A step-by-step derivation of this analytical form can be found in [12].

Note that in (7) the terms $\boldsymbol{\mu}_{ss} + \mathbf{v}_{ss}(k)$ can be regarded as the nonlinear prediction residual which is random and are dependent on the subsegment. This differs from the VTR input to the nonlinear function, which is dependent on a segment instead of on a subsegment. The finer subsegmental modeling of the prediction residual is based on our empirical observation that the accuracy of the nonlinear prediction for real speech data typically varies systematically within a phone segment. This is especially true for nonstationary phones such as stop consonants, and is less so for vowels.

IV. Results on Model Prediction for Resonance Dynamics and Reduction

In this section, we present the model simulation results, demonstrating contextually assimilated reduction. We further compare these results with the corresponding results from direct measurements of reduction in the acoustic-phonetic literature.

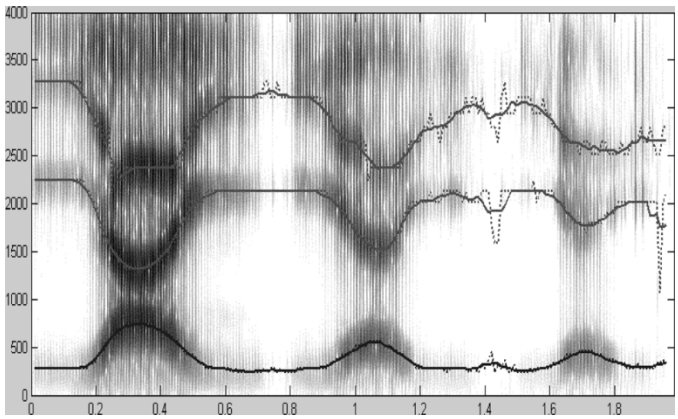

Fig. 2. Spectrogram of three renditions of /iy aa iy/ by one author, with an increasingly higher speaking rate and increasingly lower speaking efforts. The horizontal label is time, and the vertical one is frequency.

To illustrate VTR frequency or formant undershooting, we first show the spectrogram of three renditions of a three-segment /iy aa iy/ (uttered by the lead author of this paper) in Fig. 2. From left to right, the speaking rate increases and speaking effort decreases, with the durations of the /aa/'s decreasing from approximately 230 ms to 130 ms. Formant target undershooting for f_1 and f_2 is clearly visible in the spectrogram, where automatically tracked formants (using the technique described in [12]) are superimposed (as the solid lines in Fig. 2) to aid identification of the formant trajectories. (The dashed lines are the initial estimates, which are then refined to give the solid lines.)

A. Effects of Stiffness Parameter on Reduction

The same kind of target undershooting for f_1 and f_2 as in Fig. 2 is exhibited in the model prediction, shown in Fig. 3, where we also illustrate the effects of the FIR filter's stiffness parameter on the magnitude of formant undershooting or reduction. The model prediction is the FIR filter's output for f_1 and f_2 according to $\mathbf{g}(k)$ in (6). Fig. 3(a)–(c) corresponds to the use of the stiffness parameter value (the same for each formant vector component) set at $\gamma = 0.85$, 0.75 and 0.65, respectively, where in each plot the slower /iy aa iy/ sounds (with the duration of /aa/ set at 230 ms or 23 frames) are followed by the faster /iy aa iy/ sounds (with the duration of /aa/ set at 130 ms or 13 frames). f_1 and f_2 targets for /iy/ and /aa/ are set appropriately in the model also. Comparing the three plots, we have the model's quantitative prediction for the magnitude of reduction in the faster /aa/ that is decreasing as the γ value decreases.

In Fig. 4(a)–(c), we show the same model prediction as in Fig. 3 but for different sounds /iy eh iy/, where the targets for /eh/ are much closer to those of the adjacent sound /iy/ than in the previous case for /aa/. As such, the absolute amount of reduction becomes smaller. However, the same effect of the filter parameter's value on the size of reduction is shown as for the previous sounds /iy aa iy/.

B. Effects of Speaking Rate on Reduction

In Fig. 5, we show the effects of speaking rate, measured as the inverse of the sound segment's duration, on the magnitude

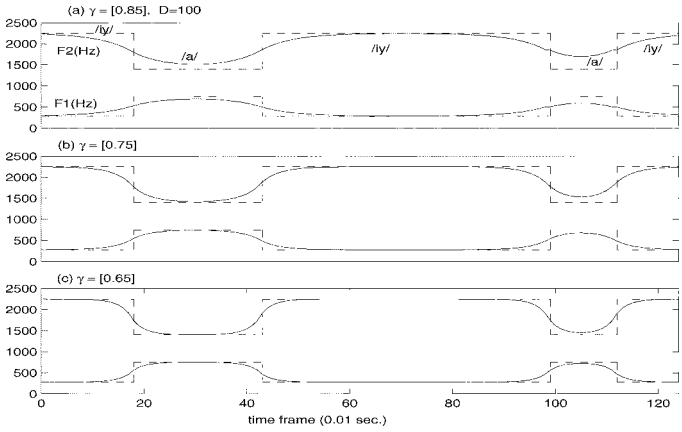

Fig. 3. f_1 and f_2 formant or VTR frequency trajectories produced from the model ($g(k)$ in (6)) for a slow /iy aa iy/ followed by a fast /iy aa iy/. (a)–(c) correspond to the use of the stiffness parameter values of $\gamma = 0.85, 0.75$, and 0.65, respectively. The amount of formant undershooting or reduction during the fast /aa/ is decreasing as the γ value decreases. The dashed lines indicate the formant target values and their switch at the segment boundaries.

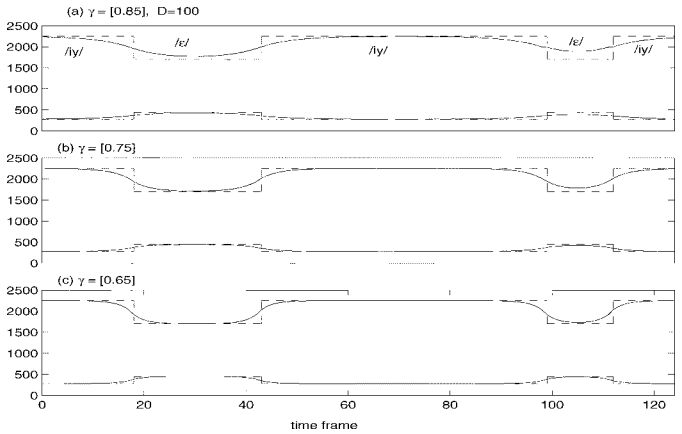

Fig. 4. Same as Fig. 3 except for the /iy eh iy/ sounds. Note that the f_1 and f_2 target values for /eh/ are closer to /iy/ than those for /aa/.

of formant undershooting. Subplots (a), (b), and (c) correspond to three decreasing durations of the sound /aa/ in the /iy aa iy/ sound sequence. They illustrate an increasing amount of the reduction with the decreasing duration or increasing speaking rate. Symbol "x" in Fig. 5 indicates the f_1 and f_2 formant values at the central portions of vowels /aa/, which are predicted from the model and are used to quantify the magnitude of reduction. These values (separately for f_1 and f_2) for /aa/ are plotted against the inversed duration in Fig. 6, together with the corresponding values for /eh/ (i.e., IPA ϵ) in the /iy eh iy/ sound sequence. The most interesting observation is that as the speaking rate increases, the distinction between vowels /aa/ and /eh/ gradually diminishes if their static formant values extracted from the dynamic patterns are used as the sole measure for the difference between the sounds. We refer to this phenomenon as "static" sound confusion induced by increased speaking rate (or/and by a greater degree of sloppiness in speaking).

C. Comparisons With Formant Measurement Data

The "static" sound confusion between /aa/ and /eh/ quantitatively predicted by the model as shown in Fig. 6 is consistent with the formant measurement data published in [25], where thousands of natural sound tokens were used to investigate the

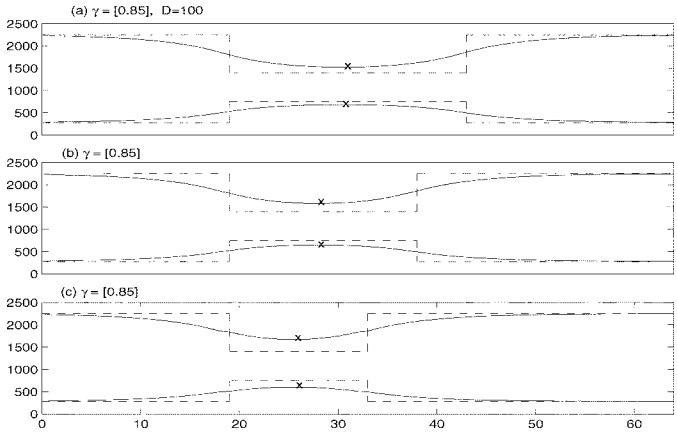

Fig. 5. f_1 and f_2 formant trajectories produced from the model for three different durations of /aa/ in the /iy aa iy/ sounds. (a) 25 frames (250 ms). (b) 20 frames. (c) 15 frames. The same γ value of 0.85 is used. The amount of target undershooting increases as the duration is shortened or the speaking rate is increased. Symbol "x" indicates the f_1 and f_2 formant values at the central portions of vowels of /aa/.

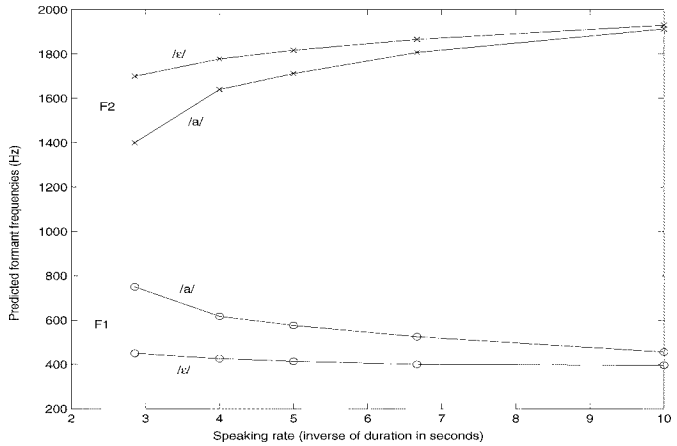

Fig. 6. Relationship, based on model prediction, between the f_1 and f_2 formant values at the central portions of vowels and the speaking rate. Vowel /aa/ is in the carry-phrase /iy aa iy/, and vowel /eh/ in /iy eh iy/. Note that as the speaking rate increases, the distinction between vowels /aa/ and /eh/ measured by the difference between their static formant values gradually diminishes. The same γ value of 0.9 is used in generating all points in the figure.

relationship between the degree of formant undershooting and speaking rate.[1] We reorganized and replotted the raw data from [25] in Fig. 7, in the same formant as Fig. 6. While the measures of speaking rate differ between the measurement data and model prediction and cannot be easily converted to each other, they are generally consistent with each other.[2] The similar trend for the greater degree of "static" sound confusion as speaking rate increases is shown clearly from both the measurement data (Fig. 7) and prediction (Fig. 6).

D. Model Prediction of VTR Trajectories for Real Speech Utterances

We have used Stage I of the model to predict actual VTR frequency trajectories for speech utterances from TIMIT database. Only the phone identities and their boundaries are input to the

[1]We are grateful to Dr. M. Pitermann for providing us with raw data of formant measurements published in [25], which allows us to do the replotting.

[2]We again thank Dr. M. Pitermann for useful discussions on this point.

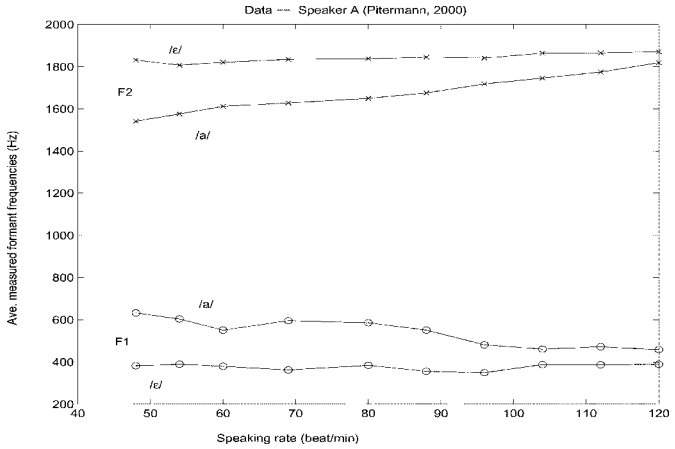

Fig. 7. Formant measurement data from literature are reorganized and plotted, showing similar trends to the model prediction under similar conditions.

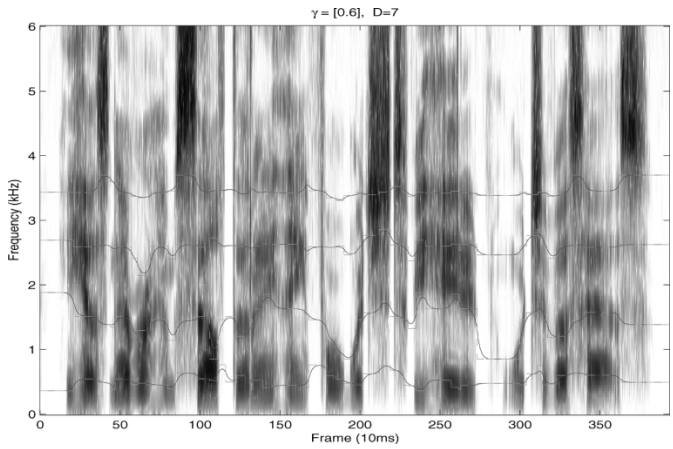

Fig. 8. $f_1/f_2/f_3/f_4$ VTR frequency trajectories (smooth lines) generated from the FIR model (Stage I) using the phone sequence and duration of a speech utterance (SI1039) taken from the TIMIT database. The target sequence is shown as stepwise lines, switching at the phone boundaries labeled in the database. They are superimposed on the utterance's spectrogram. The utterance is "*He has never, himself, done anything for which to be hated – which of us has*. "

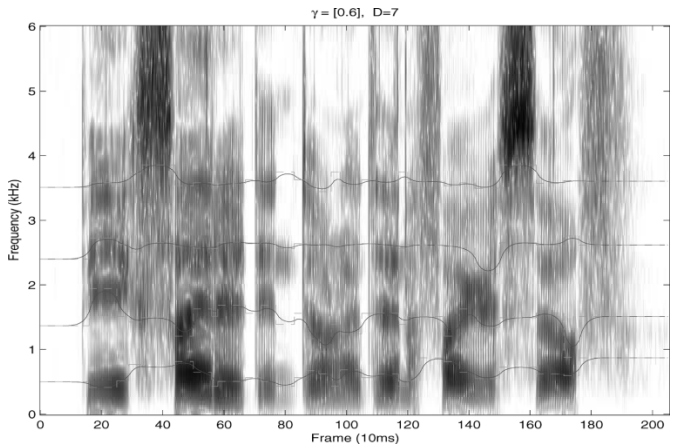

Fig. 9. Same as Fig. 8 except with another utterance "*Be excited and don't identify yourself*" (SI1669).

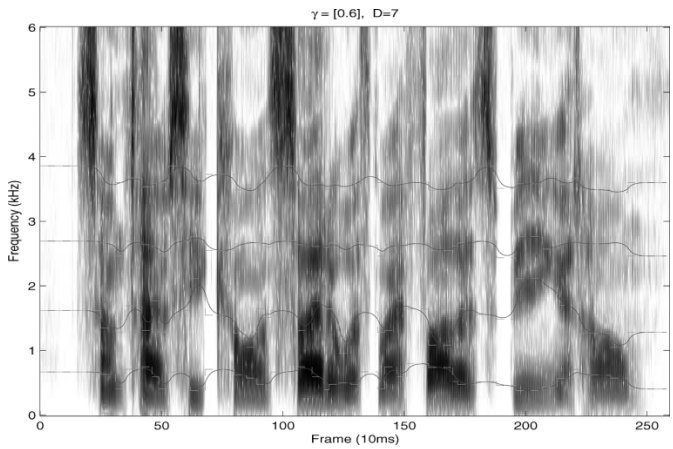

Fig. 10. Same as Fig. 8 except with the third utterance "*Sometimes, he coincided with my father's being at home*" (SI2299).

model for the prediction, and no use is made of speech acoustics; i.e., only Stage I of the model, and not Stage II, is used. This differs from the task of formant or VTR tracking where speech acoustics is always used [12].

Given the phone sequence in any utterance, we first break up the compound phones (affricates and diphthongs) into their constituents. Then we obtain the initial VTR target values based on limited context dependency by table lookup (see details in [9, Ch. 13]). Then automatic and iterative target adaptation is performed for each phone-like unit based on the difference between the results of a VTR tracker (described in [11]) and the VTR prediction from the FIR filter model. (This iterative adaptation algorithm will not be described in this paper due to space limitation.) Note these target values are provided not only to vowels, but also to consonants for which the resonance frequency targets are used with weak or no acoustic manifestation. The converged target values, together with the phone boundaries provided from the TIMIT database, form the input to the FIR filter in Stage I of the model and the output of the filter gives the predicted VTR frequency trajectories.

Three example utterances from TIMIT (SI1039, SI1669, and SI2299) are shown in Figs. 8–10. The step-wise dashed lines ($f_1/f_2/f_3/f_4$) are the target sequences as inputs to the FIR filter, and the continuous lines ($f_1/f_2/f_3/f_4$) are the outputs of the filter as the predicted VTR frequency trajectories. Parameters γ and D are fixed and not automatically learned. To facilitate assessment of the accuracy in the prediction, the inputs and outputs are superimposed on the spectrograms of these utterances, where the true resonances are shown as the dark bands. For the majority of frames, the filter's output either coincides or is close to the true VTR frequencies, even though no acoustic information is used. Also, comparing the input and output of the filter, we observe only a rather mild degree of target undershooting or reduction in these and many other TIMIT utterances we have examined but not shown here.

V. Results on Model Prediction for Cepstrum Dynamics

The predicted VTR dynamics by model Stage I in Figs. 8–10 are fed into model Stage II, to produce the predicted LPC cepstra in Figs. 11–13, respectively, for the previous three

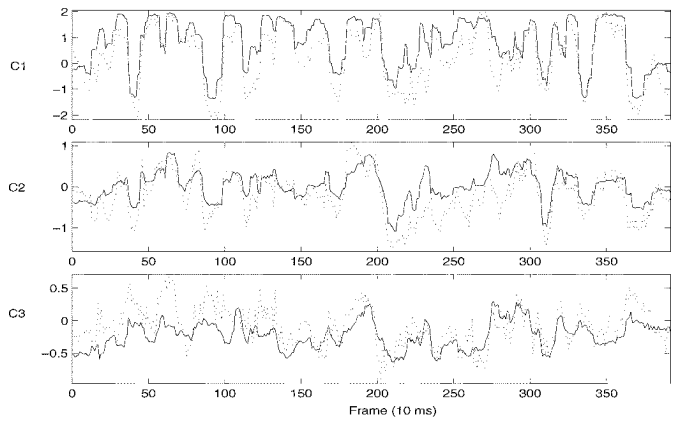

Fig. 11. LPC cepstra with order one (C1), two (C2), and three (C3) predicted from the Stage II of the model (solid lines) using the input from the FIR model's output for utterance SI1039. Dashed lines are the LPC cepstral data C1, C2, and C3 computed directly from the waveform.

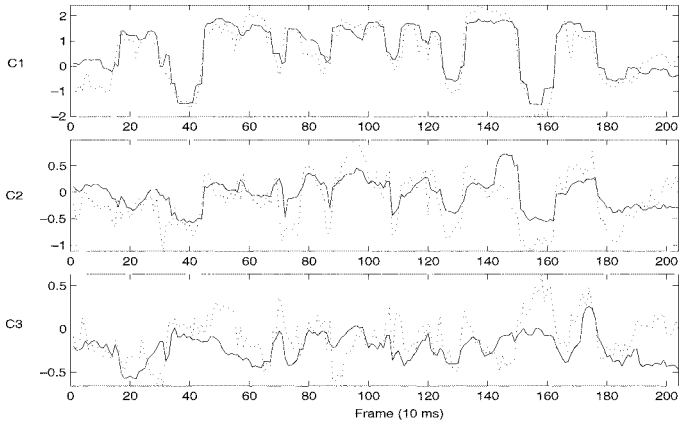

Fig. 12. Same as Fig. 11 except with the second utterance (SI2299).

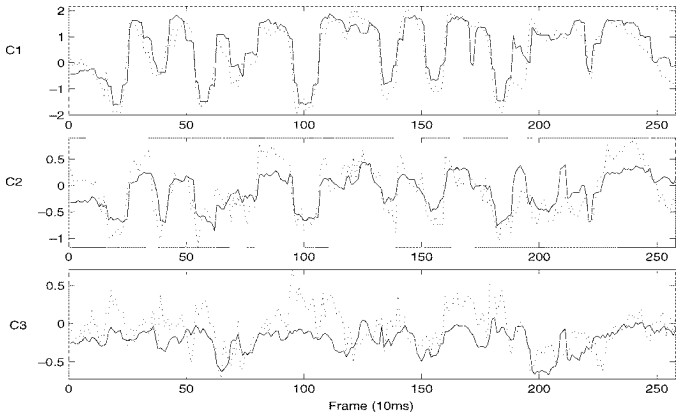

Fig. 13. Same as Fig. 11 except with the third utterance (SI1669).

example TIMIT utterances. Note that the model prediction includes residual means, which are trained from the full TIMIT data set using an hidden Markov model toolkit (HTK) tool. The zero-mean random component of the residual is ignored in these figures. The residual means for the substates (three for each phone) are added sequentially to the output of the nonlinear function (8), assuming each substate occupies three equal-length sub-segments of the entire phone segment length provided by TIMIT database. To avoid display cluttering,

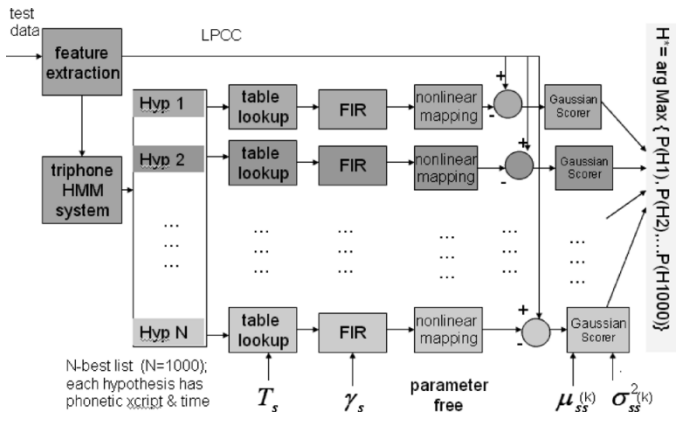

Fig. 14. Block diagram for the N-best evaluation procedure.

only LPC cepstra with orders one (C1), two (C2), and three (C3) are shown here, as the solid lines. Dashed lines are the LPC cepstral data C1, C2, and C3 computed directly from the waveforms of the same utterances for comparison purposes. The data and the model prediction generally agree with each other, somewhat better for lower order cepstra than for higher order ones. We found that these discrepancies are generally within the variances of the prediction residuals automatically trained from the entire TIMIT training set (using an HTK tool for monophone HMM training).

VI. Applications to Phonetic Recognition

A. Recognizer Design

In the two-stage implementation of the coarticulation and reduction model presented so far, we ignore the variability in the VTR dynamics in the prediction of the cepstrum dynamics. This significantly simplifies the application of the model as a phonetic recognizer. That is, given any phone sequence with possible phone segmentation (e.g., derived from N-best hypotheses), model Stage I generates deterministic VTR trajectories. Feeding these into probabilistic model Stage II, a likelihood can be computed using the Gaussian assumption of the cepstral residual. This scoring mechanism allows the recognizer to perform N-best rescoring in a straightforward manner. The block diagram for the recognizer that executes the N-best evaluation procedure is shown in Fig. 14, where Stage-I and Stage-II of the model for each of the N-best hypotheses are represented by the blocks labeled as "FIR (filter)" and "nonlinear mapping," respectively. The "table lookup" block represents the construction process for forming the VTR target sequence, using the target values stored in the table that are trained in advance.

In the recognizer evaluation procedure, shown as the operation following the "nonlinear mapping" block in Fig. 14, the nonlinear prediction of LPC cepstra according to (8) is directly subtracted from the LPC cepstral data as the recognizer's input acoustic features. This difference, separately for each of the N-best hypotheses computed from a state-of-the-art triphone HMM, forms the residual sequence that follows a monophone Gaussian HMM. We use an HTK tool (Hvite) to directly compute the likelihood for the residual sequence, which is exactly the same as the likelihood for the original LPC

cepstral data sequence, given each of the N-best hypotheses. This likelihood computation operation is shown in the blocks labeled as "Gaussian score" in Fig. 14. The results of this set of computations are reranked as the recognizer's final output for the recognition accuracy determination, which we will present shortly.

One specific aspect of the recognizer design that we have developed in this work is to naturally incorporate the delta and acceleration features into the recognizer. We first decompose the LPC cepstral feature differentials (delta and acceleration) for each frame in the data into the part that can be predicted from the VTR [according to (8)] and the part that cannot be predicted [i.e., the residual terms $\boldsymbol{\mu}_{ss} + \mathbf{v}_{ss}(k)$ in (7)]. Thus, the basic Stage-II model as described in (7) is expanded to ones that consist of delta and acceleration components as well. For the predictable part in these new components, we directly compute the frame differentials of the predicted LPC cepstral values from (8). For the unpredictable part that cannot make use of any information from the model, we train the delta and acceleration parameters of means and variances for the residuals using the corresponding frame-differential LPC cepstral training data. One desirable property of this technique for treating delta and acceleration features is that in the degenerative case where the predictive model component (8) is removed by setting it to zero, the recognizer automatically becomes a conventional (monophone) HMM system.

B. Recognizer Training

To compute the previous residual likelihood requires that residual means and variances of each substate of each phone in the N-best hypotheses be known. These (monophone) parameters are trained automatically from the TIMIT training set. Again, given the training script, including both phone sequences and phone boundaries, model Stage I generates deterministic VTR trajectories and model Stage II generates predicted cepstral trajectories. Subtracting the predicted cepstral trajectories from the cepstral training data on a frame-by-frame basis gives residuals for the training set. Treating these residuals as the "training data," we apply an HTK tool to train a set of residual monophone HMMs. The mean and variance parameters of these models are used for N-best rescoring as described in the preceding subsection.

Because our current simplistic two-stage implementation of the model ignores the VTR variability across speakers and across utterances, it is necessary to provide reasonably accurate VTR targets in order to obtain a high likelihood for the correct phone sequence. To achieve this, we have developed and applied an iterative target training and adaptation technique for each of the N-best hypotheses before the rescoring process takes place as described previously.

Finally, the parameters in model Stage I, D and γ are empirically set for the TIMIT experiments. They are determined by fitting the model prediction to the formant data in training utterances. It is found that the fixed values of $D = 7$ and $\gamma = 0.6$ already provide good fit to the data for most of the TIMIT training data we have examined. It appears that these parameters may not need to be made as dependent on phones, on speakers, or utterances for the TIMIT data.[3]

C. Phonetic Recognition Task and Results

The phonetic recognition experiments which we carried out to evaluate our two-stage coarticulatory model are based on the widely used TIMIT database. We built two-stage acoustic models using the standard 61 label set, which are folded into 48 classes, in training the residual means and variances for each subsegment of each class. VTR targets are trained and then adapted for each phone segment instead of subsegment. For diphthongs and affricates, two separate targets are trained, assuming one target following another. Phonetic recognition errors are tabulated using the 39 labels adopted by many other researchers to report recognition results. Model parameters are trained on the designated training set of 462 speakers, and results are reported on the standard core test set with a total of 192 utterances by 24 speakers.[4]

We use the N-best rescoring paradigm to evaluate the new two-stage coarticulatory model. For each of the 192 core test utterances, we use a standard triphone HMM with a decision tree to generate a very large N-best list where $N = 2000$. A biphone language model is used to generate this N-best list in order to improve the quality of the list as much as possible. Also, Mel-frequency cepstral coefficients with delta and acceleration features are used in generating this N-best list. The reason why a language model and Mel-frequency warping are used for the HMM to generate the N-best list is because we desire to create the list with the highest quality possible in order to provide the richest set of candidates possible for scoring the new recognizer based on the two-stage, target-filtering model. The oracle phone error rate is about 17% for the full top 2000 list. Although the use of Mel-frequency warping for cepstral features is known to benefit the HMM performance, it has not been used for the two-stage, target-filtering model. This is because of the requirement for model Stage-II to generate not the LPC cepstrum as in (8) but its Mel-warped version. No simple analytical form of the mapping function is available and to predict the Mel-warped cepstra requires more complicated model Stage-II than the one presented in this paper.

With the use of a flat phone language model and of the LPC cepstra (including delta and acceleration) as features, the phone recognition accuracy for the standard triphone HMM is 64%, as shown at the top row in Table I. This baseline result is produced by a full decoder in HTK. Under the above experimental conditions, we evaluate the top-one accuracy (100% minus substitution, deletion, and insertion errors) using the $N = 2000$ best list for our new coarticulatory model. It gives 71.91%, significantly higher than the triphone HMM (22% relative error rate reduction), despite the use of only context independent model parameters. To assess the effect of using the N-best list generated by the HMM which is substantially different from the new model, i.e., the effect of combining different recognition

[3]This is not the case, however, for other speech data such as Switchboard that we have examined.

[4]We thank Dr. J. Glass of MIT who prompted us to use the core test set, and provided us with the file list, for the evaluation.

TABLE I
PHONETIC RECOGNITION ACCURACY ON TIMIT CORE TEST SET (192 UTTERANCES) USING THE BI-DIRECTIONAL TARGET-FILTERING MODEL OF SPEECH COARTICULATION WITH A TWO-STAGE IMPLEMENTATION (LABELED AS "NEW MODEL") IN COMPARISON WITH A CONVENTIONAL RECOGNIZER (LABELED AS "TRIPHONE-HMM"). NO LANGUAGE MODEL IS USED, AND THE FEATURES FOR BOTH SYSTEMS ARE THE SAME LPC CEPSTRA. THE HMMS MEAN AND VARIANCE PARAMETERS ARE CONDITIONED ON TRIPHONES CLUSTERED BY DECISION TREE, USING THE STANDARD HTK TOOLS. THE NEW MODEL IS PARAMETERIZED BY CONTEXT-INDEPENDENT, SEGMENT-SPECIFIC VTR TARGET VECTORS, AND BY THE CONTEXT-INDEPENDENT, SUBSEGMENT-SPECIFIC RESIDUAL MEANS AND VARIANCES. N-BEST RESCORING IS USED FOR THE NEW MODEL, WITH WIDELY VARYING SIZES OF N TO ASSESS THE ROVER EFFECT

Model Type	N-in-Nbest	Top-one Accuracy
Triphone-HMM	(full decoding)	64.00%
New Model	2000	71.91%
	1000	71.80%
	800	72.18%
	500	72.36%
	300	72.47%
	100	72.40%

systems (as related to the ROVER effect[5]), we rescore our new model using a varying size N in the N-best list. The top-one accuracies are listed in the remaining rows in Table I. As can be seen, the ROVER effect is relatively minor, and it is virtually eliminated by using large N-best lists for N being between 1000 and 2000. The converging accuracy of 71.91% is, thus, established that is not biased by the ROVER effect. Note that the previous results are obtained with no use of combination between the original HMM scores and the new model's scores. The new model's scores are used alone to do reordering of the N-best list

VII. SUMMARY AND CONCLUSION

In this paper, we first presented a quantitative model for predicting VTR dynamics, accounting for the related reduction and "static" speech sound confusion phenomena. This model is based on bidirectional filtering of phone-dependent, VTR target sequences implemented with a temporally symmetric FIR digital filter. This forms Stage I of an overall two-stage speech generation model, where the final Stage II takes the output of Stage I as its input and generates the LPC cepstra via a parameter-free, analytical nonlinear prediction function. The errors of this nonlinear prediction for the LPC cepstral speech data are represented by phone-subsegment dependent Gaussian random variables, whose parameters are automatically trained from a set of phonetically labeled training data.

We present details of model simulation that demonstrates quantitative effects of speaking rate and segment duration on the magnitude of reduction. Both VTR dynamics and cepstral dynamics as outputs from model Stage I and Stage II, respectively, are compared with and shown to be close to real speech data.

[5]The strict ROVER effect refers to that for combining system outputs [13]. Here we have a slightly different condition of combining two different system properties at an intermediate level.

A phonetic recognizer is constructed using this new generative model of speech dynamics, and is evaluated in the standard TIMIT phonetic recognition task. N-best rescoring is used for the evaluation, with varying size of N from 100 to 2000. We demonstrate 22% error rate reduction using the new model compared with the standard HMM under the following three identical conditions: 1) the same input feature parameters of LPC cepstra to the recognizers; 2) the same full set of TIMIT training data; and 3) the same flat language models. This significant performance gain is validated after removing the ROVER effect by using an increasingly larger size of N-best lists where a converging recognition accuracy is observed.

The development of the model presented in this paper is motivated by phonetic theories and experiments on sound reduction in free-style speech. We intend to use the model as one major source of *a priori* knowledge about the speech structure for automatic recognition of conversational speech. We have accumulated evidence that the strong reduction and "static" sound confusion in this mixed style of speech, ranging widely in the hyper-hypo speaking continuum, are responsible for many recognition errors by state-of-the-art automatic systems. The new model is demonstrated in simulation experiments to be capable of resolving the confusion with dynamic speech specification, thus, it would be more useful for conversational speech. conversational speech recognition. Our future research in this direction involves relaxing the current simplifying assumption of deterministic VTR dynamics at Stage I of the model, aiming at an integrated solution that simultaneously takes into account the inevitable variabilities in both hidden VTR and observed acoustic domains. We are currently also working on extending the LPC cepstral features to the Mel-warped features within the same generative modeling framework as presented in this paper.

ACKNOWLEDGMENT

The authors wouldl like to thank Dr. M. Pitermann for providing them with the raw data of formant measurements and for useful discussions. They would also like to thank Dr. M. Seltzer who built the triphone HMM system for them to generate a high-quality N-best list for rescoring of the new model.

REFERENCES

[1] R. Bakis, "Coarticulation modeling with continuous-state HMMs," in *Proc. IEEE Workshop Automatic Speech Recognition*, New York, 1991, pp. 20–21.

[2] J. Bilmes, "Graphical models and automatic speech recognition," in *Mathematical Foundations of Speech and Language Processing*, M. Johnson, M. Ostendorf, S. Khudanpur, and R. Rosenfeld, Eds. New York: Springer-Verlag, 2004, pp. 135–186.

[3] J. Bridle *et al.*, "An investigation of segmental hidden dynamic models of speech coarticulation for automatic speech recognition," in *Proc. Final Report Workshop on Language Engineering, Center for Language and Speech Processing at The Johns Hopkins University*, 1998, pp. 1–61.

[4] C. Chelba and F. Jelinek, "Structured language modeling," *Comput. Speech Lang.*, pp. 283–332, Oct. 2000.

[5] L. Deng, "A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal," *Signal Process.*, vol. 27, pp. 65–78, 1992.

[6] ——, "A dynamic, feature-based approach to the interface between phonology and phonetics for speech modeling and recognition," *Speech Commun.*, vol. 24, no. 4, pp. 299–323, 1998.

[7] ——, "Computational models for speech production," in *Computational Models of Speech Pattern Processing*, K. Ponting, Ed. Berlin, Germany: Springer-Verlag, 1999, pp. 199–213.
[8] ——, "Switching dynamic system models for speech articulation and acoustics," in *Mathematical Foundations of Speech and Language Processing*, M. Johnson, M. Ostendorf, S. Khudanpur, and R. Rosenfeld, Eds. New York: Springer-Verlag, 2004, pp. 115–134.
[9] L. Deng and D. O'Shaughnessy, *SPEECH PROCESSING — A Dynamic and Optimization-Oriented Approach*. New York: Marcel Dekker, 2003.
[10] L. Deng and D. Braam, "Context-dependent Markov model structured by locus equations: Applications to phonetic classification," *J. Acoust. Soc. Amer.*, vol. 96, pp. 2008–2025, Oct. 1994.
[11] L. Deng, L. Lee, H. Attias, and A. Acero, "A structured speech model with continuous hidden dynamics and prediction-residual training for tracking vocal tract resonances," in *Proc. IEEE ICASSP*, vol. I, May 2004, pp. 557–560.
[12] L. Deng, A. Acero, and I. Bazzi, "Tracking vocal tract resonances using a quantized nonlinear function embedded in a temporal constraint," IEEE Trans. Speech Audio Process., to be published.
[13] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proc. Automatic Speech Recognition and Understanding*, 1997, pp. 347–354.
[14] Y. Gao, R. Bakis, J. Huang, and B. Zhang, "Multistage coarticulation model combining articulatory, formant and cepstral features," in *Proc. ICSLP*, vol. 1, 2000, pp. 25–28.
[15] T. Gay, "Effect of speaking rate on vowel formant movements," *J. Acoust. Soc. Amer.*, vol. 63, pp. 223–230, 1978.
[16] M. Siu, R. Iyer, H. Gish, and C. Quillen, "Parametric trajectory mixtures for LVCSR," in *Proc. ICSLP*, Sydney, Australia, 1998, pp. 3269–3272.
[17] S. Hertz, "Streams, phones, and transitions: Toward a new phonological and phonetic model of formant timing," *J. Phonet.*, vol. 19, pp. 91–109, 1991.
[18] B. Lindblom, "Spectrographic study of vowel reduction," *J. Acoust. Soc. Amer.*, vol. 35, pp. 1773–1781, 1963.
[19] W. Holmes and M. Russell, "Probabilistic-trajectory segmental HMMs," *Comput. Speech Lang.*, vol. 13, pp. 3–37, 1999.
[20] D. Klatt, "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Amer.*, vol. 99, no. 3, pp. 971–995, 1980.
[21] B. Lindblom, "Explaining phonetic variation: A sketch of the H & H theory," in *Speech Production and Speech Modeling*, W. Hardcastle and A. Marchal, Eds. Norwell, MA: Kluwer, 1990, pp. 403–439.
[22] J. Ma and L. Deng, "Efficient decoding strategies for conversational speech recognition using a constrained nonlinear state-space model for vocal-tract-resonance dynamics," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 590–602, Nov. 2003.
[23] S. Moon and B. Lindblom, "Interaction between duration, context, and speaking style in English stressed vowels," *J. Acoust. Soc. Amer.*, vol. 96, pp. 40–55, 1994.
[24] M. Ostendorf, V. Digalakis, and J. Rohlicek, "From HMM's to segment models: A unified view of stochastic modeling for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 5, pp. 360–378, Sep. 1996.
[25] M. Pitermann, "Effect of speaking rate and contrastive stress on formant dynamics and vowel perception," *J. Acoust. Soc. Amer.*, vol. 107, pp. 3425–3437, 2000.
[26] L. Pols, "Psycho-acoustics and speech perception," in *Computational Models of Speech Pattern Processing*, K. Ponting, Ed. Berlin, Germany: Springer-Verlag, pp. 10–17.
[27] R. Rose, J. Schroeter, and M. Sondhi, "The potential role of speech production models in automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 99, pp. 1699–1709, 1996.
[28] K. Stevens, "On the quantal nature of speech," *J. Phonet.*, vol. 17, pp. 3–45, 1989.
[29] J. Sun and L. Deng, "An overlapping-feature based phonological model incorporating linguistic constraints: Applications to speech recognition," *J. Acoust. Soc. Amer.*, vol. 111, no. 2, pp. 1086–1101, Feb. 2002.
[30] D. van Bergem, "Acoustic vowel reduction as a function of sentence accent, word stress and word class," *Speech Commun.*, vol. 12, pp. 1–12, 1993.
[31] W. Wang, A. Stolcke, and M. Harper, "The use of a linguistically motivated language model in conversational speech recognition," in *Proc. IEEE ICASSP*, vol. I, May 2004, pp. 261–264.
[32] J. Wouters and M. Macon, "Control of spectral dynamics in concatenative speech synthesis," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 1, pp. 30–38, Jan. 2001.
[33] J. Zhou, F. Seide, and L. Deng, "Coarticulation modeling by embedding a target-directed hidden trajectory model into HMM," in *Proc. IEEE ICASSP*, vol. I, Apr. 2003, pp. 744–747.

Li Deng, photograph and biography not available at the time of publication.

Dong Yu, photograph and biography not available at the time of publication.

Alex Acero, photograph and biography not available at the time of publication.