Handling Phonetic Context and Speaker Variation in a Structure-Based Speech Recognizer

Dong Yu, Li Deng, and Alex Acero

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

{dongyu, deng, alexac}@microsoft.com

Abstract

Recently we have developed a novel type of structure-based speech recognizer, which uses parameterized, non-recursive "hidden" trajectory model of vocal tract resonances (VTR) or formants to capture the dynamic structure of long-range speech coarticulation and reduction. The underlying model of this recognizer carries out bi-directional FIR filtering on the piecewise constant sequences of the VTR targets. In this paper, we elaborate on two key aspects of the model. First, the phonetic context controls the movement direction and thus the formation of the VTR trajectories. This provides "structured" context dependency for speech acoustics without using context dependent parameters as required by HMMs. Second, VTR targets as the key context-independent parameters of the model vary across speakers. We describe an effective target-value normalization algorithm that can be applied to both training and unknown test speakers. We report experimental results demonstrating the effectiveness of the normalization algorithm in the context of structure-based speech recognition. We also provide computational analysis on the HTM-based speech decoder.

Index Terms: hidden trajectory model, phonetic contexts, normalization, vocal tract resonance, targets

1. Introduction

We recently developed a version of structure-based speech recognizer, which we call the hidden trajectory model (HTM), where a non-recursive, parametric form of the time-series model is used to functionally represent the dynamic structure of speech articulation in the domain of vocal tract resonance (VTR). Various aspects of the model development including its training and decoding algorithms as well as experimental results have been described in [1][2][12][13][6]. HTM can be considered as a special and highly structured member in the family of generic stochastic segment models [8]. In this model, dynamic structure of speech is represented in the unobserved VTR domain to characterize long-span contextual influence among phonetic units in fluent speech utterances. One key idea is the use of bi-directional and parametric target filtering to model speech coarticulation and context-assimilated phonetic reduction.

In a series of earlier papers, we described in detail the formulation of the bi-directional filter [1], the cepstral-residual parameter learning algorithm [2], the VTR target learning algorithm [13], and synchronous and asynchronous decoders for the HTM [12][6]. In this paper, we elaborate on two additional aspects in HTM: 1) the role of phonetic contexts and its incorporation in speech recognition algorithms, and 2) the way to handle speaker variation via the normalization of resonance target parameters in HTM. These aspects have been discussed yet not emphasized in our earlier papers [12][13].

The phonetic contexts control the values of the VTR trajectory. Proper incorporation of phonetic contexts is a key aspect in constructing accurate models for speech dynamics. In HTM, phonetic contexts are partially modeled by the bi-directional target filter and partially modeled by the proper selection of the context-sensitive "HTM-unit" sets. In this paper, we draw attention to the importance of the correct selection of the HTM-unit set. We explain why the commonly used phone set in HMM systems is not sufficient to accurately model the contextual relationships embodied in HTM, and show the construction of the HTM-unit set together with the motivations.

It has been traditionally held that more complex statistical models often would have a harder time to incorporate detailed knowledge into the associated algorithms. As an example, in relatively simple models such as HMMs, speaker variation can be straightforwardly handled by pooling all data from many speakers in training. But for more complex models such as HTMs, such pooling would not work since some key parameter set (i.e., VTR targets) in the model are inherently speaker specific. While statistical distributions can be used to represent the randomness of the VTR targets due to speaker variation, this would significantly increase phonetic confusability. Special normalization techniques have been developed, which will be described in detail in this paper.

The organization of this paper is as follows. In Section 2, we illustrate the idea of trajectory estimation using the bi-directional target filters in the hidden trajectory model and the incorporation of the phonetic contexts. In Section 3, we show the importance of normalizing the targets and describe two methods for target normalization and prediction. Experimental results are shown in Section 4, and computational analysis shown in Section 5.

2. HTM and Phonetic Contexts

In HTM, phonetic context controls the movement direction of the VTR trajectories and subsequently their full formation when initial conditions are given. This control provides structured context dependency for speech acoustics without using context dependent parameters as required by HMMs. We discuss in this section details on how the phonetic context is incorporated in the HTM.

As a generative model, the HTM first converts the temporally segmented, piecewise constant VTR target sequence (with sharp jumps at the segment boundaries) into smooth VTR trajectories using a bi-directional filter given the phone-sequence hypothesis and the boundaries [1][2]. The obtained VTR trajectories are next converted, via a nonlinear function, into the cepstral trajectories with sub-unit dependent bias parameters. In the decoding/recognition process, the above "generated" cepstral trajectories are "compared" with the measured trajectories as the input data that are directly computed from the audio signal -- this comparison results in

the likelihood for each of the possible hypotheses. The hypothesis with the highest likelihood is chosen as the results of the recognition using the HTM.

Figure 1 provides an illustration of using the bi-directional filter to smooth the segmental VTR target sequence. Each HTM-unit (phone-like) is associated with a unique target vector (three dimensions in this example) and timing. Note that the filtered VTR trajectories exhibit both forward and backward coarticulation, since the VTR value at each time depends on not only the current unit's VTR target value but also those of the adjacent units. In this way, the phonetic context information is directly incorporated into the HTM via the filtering operation. In [1], it was shown that this filtering operation can quantitatively predict the magnitude of contextually assimilated reduction and coarticulation. Below we draw attention to the importance of appropriate selection of the HTM-unit/phone set and the related phonetic context, and describe the actual selection in our implemented HTM-based recognition system.

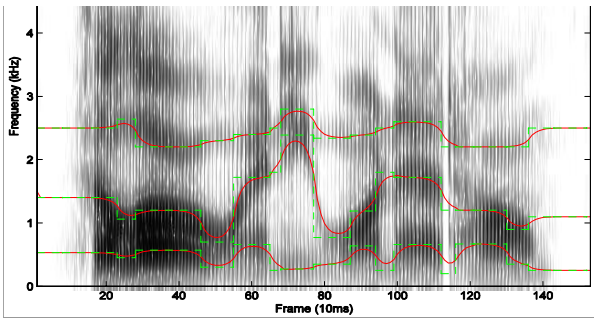

Figure 1: *An example of using the bi-directional filter on the VTR target sequence (dotted green lines) to determine the VTR trajectories (solid red lines). Three dimensions of the target and VTR are illustrated, for F1, F2, and F3, respectively. Both the target sequence and the filtered VTR results are superimposed on the spectrogram of the speech utterance.*

The HTM-units are based on the conventional phones with four main categories of modifications that take into account regular effects of certain phonetic contexts. The first category concerns the place-of-articulation context including labial, labial-dental, and velar features for American English. As the HTM-units, the labial (/b/, /p/, /m/), labial-dental (/f/, /v/), and velar (/g/, /k/, /ng/) consonants are made conditioned on whether the following phone has the 'front' feature or otherwise. That is, when these consonants are followed by one of the following front vowels: /ae/, /eh/, /ih/, /iy/, /y/, or /ey/, we name them /b_f/, /p_f/, /m_f/, /f_f/, /v_f/, /g_f/, /k_f/, and /ng_f/, respectively, as distinguished HTM-units from the non-front-context counterparts. These two different sets of HTM-units (for the same phones) have different VTR targets which are trained seperately. In the decoding process, these two sets of units are also selected based on the following units in the hypotheses.

Our second category of unit modifications (for our TIMIT-related experiments reported in this paper) concern the target-less TIMIT labels including pause, silence, and allophones /hh/ (unvoiced) or /hv/ (voiced). Figure 2 shows automatically extracted VTR/formant trajectories (F1, F2, F3) of a TIMIT utterance containing labels /hh/ and /hv/. There are no intrinsit articulatory configurations for these sounds, and hence no VTR targets associated with them. Instead, the VTR targets of the adjacent hones are 'inherited' to these phones in order to predict the VTR trajectories for them. From Figure 2, we see that the VTR trajectories for 'target-less' /hh/ and /hv/ are reasonably close to the above prediction.

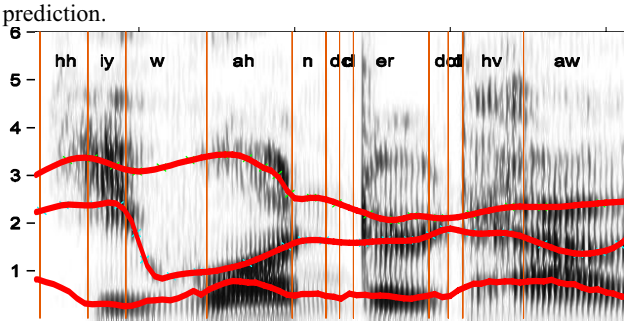

Figure 2: VTR/formant trajectories (F1, F2, F3) of a TIMIT utterance containing labels /hh/ and /hv/ that do not have intrinsic VTR targets.

The third category of HTM-unit modifications and the associated phonetic context are related to the `composite' phones including affricates (/jh/, /ch/), and diphthongs (/ey/, /aw/, /ay/, /oy/, /ow/). To simplify the modeling, we split each composite phone into its constituents. That is, there are two sequentially connected VTR targets for each composite phone (with no specification of where the target switching is).

In the fourth category of HTM-unit modifications, we use the same tying scheme as the classical one proposed in [5] for the TIMIT database. That is, the VTR targets are the same for all units within each of the following sets: {/em/, /m/}, {/en/, /nx/, /n/}, {/el/, /l/}, {/ao/, /aa/}, {/ux/, /uw/}, {/ix/, /ax-h/, /ax/}, {/axr/, /er/}, {/q/, /kcl/, /pcl/, /tcl/, /cl/}, {/bcl/, /dcl/, /gcl/, /vcl/}.

3. Target Normalization across Speakers

In this section, we present a novel algorithm for normalizing one key set of HTM parameters, HTM-unit-dependent VTR targets which are context independent, across speakers. One simplest approach would be to model all speakers with a single set of VTR targets but large variances. However, such simplicity would produce the VTR trajectories with wrong HTM-unit sequences, which may match well with data for some speakers, but not for other speakers (since VTR targets are related closely to the vocal tract length of the speaker). This problem has been analyzed in [13], where a peaker-adaptive target training algorithm was described using VTRs estimated by a high-accuracy VTR tracking technique presented in [3]. In this paper, we further developed the target-normalization algorithm with new experimental results. Specifically, we show evidences that our new target-normalization algorithm improves the performance of the VTR prediction. Importantly, the improvement is most striking when the normalization factor needs to be estimated using a small segments of the test set.

In HTM and in absence of acoustic data, the VTR trajectory is predicted using the sequence of VTR target values (corresponding to the HTM-units) and their boundaries. The prediction is carried out by filtering the VTR targets using the bi-directional FIR filters [1]. In its simplest form, we assume that the target values for unit s, $\bar{\mathbf{T}}_s$ (a vector of F1, F2, F3, and F4 in our implementation), is the same for all speakers in the dataset. The residual difference $\mathbf{r}_s$ between $\bar{\mathbf{T}}_s$ and the true target $\mathbf{T}_s$ follows the normal distribution given by $\mathbf{r}_s \sim N(\mathbf{r}_s; \mathbf{0}, \boldsymbol{\varphi}_s)$, where $\mathbf{r}_s = \mathbf{T}_s - \bar{\mathbf{T}}_s$. That is,

$$\mathbf{T}_s = \bar{\mathbf{T}}_s + N(\mathbf{r}_s; \mathbf{0}, \boldsymbol{\varphi}_s). \quad (1)$$

The variance $\boldsymbol{\varphi}_s$ above indicates the size of the residual and can be large since VTR targets depend on the vocal tract length, which varies widely among different speakers in the training data and hence is often subject to special treatment before feeding the features into automatic speech recognizers (e.g., [4][7][9][10][11][14]). The estimation accuracy of the VTR targets directly controls that of the VTR trajectories (and hence the qualities of speech recognition whose input acoustic features are direct function of the VTR trajectories). The goal of the target-normalization algorithm described in this section below is to reduce the variation of the VTR target values (and hence of the predicted VTR trajectories), and to account for realistic speaker-dependent target parameters based on the relatively invariant, normalized values.

A straight-forward approach to accounting for speaker-dependent target parameters in the HTM is to scale the "generic", speaker-normalized VTR targets $\bar{\mathbf{T}}_s$ according to

$$\bar{\mathbf{T}}_{spk,s} = \boldsymbol{\beta}_{spk} \bullet \bar{\mathbf{T}}_s, \quad (2)$$

where $\boldsymbol{\beta}_{spk}$ is the speaker-dependent normalization or scaling factor inversely proportional to the vocal tract length of the speaker spk. The dot operation above denotes an element-by-element multiplication. This approach is based on the assumption that the ratio of the average VTR values (either in terms of targets or of actual values in the trajectories) of two speakers is a good estimate of a fixed physical property (the ratio of two speakers' vocal tract lengths). That is,

$$\frac{\bar{\mathbf{T}}_{spk,s}}{\bar{\mathbf{T}}_s} = \frac{\bar{\mathbf{z}}_{spk,s}}{\bar{\mathbf{z}}_s} = \boldsymbol{\beta}_{spk}. \quad (3)$$

This assumption has been commonly used (e.g., [4][7][9][10][11][14]). Now with target normalization, a speaker-specific target will become

$$\mathbf{T}_{spk,s} = \boldsymbol{\beta}_{spk} \bullet \bar{\mathbf{T}}_s + N(\mathbf{r}_{spk,s}; \mathbf{0}, \boldsymbol{\varphi}_{spk,s}). \quad (4)$$

The goal then is to estimate $\boldsymbol{\beta}_{spk}$ so that $\boldsymbol{\varphi}_{s,spk} < \boldsymbol{\varphi}_s$.

3.1. Method one

As developed and reported in [13], the most intuitive method of estimating $\boldsymbol{\beta}_{spk}$ is

$$\bar{\boldsymbol{\beta}}_{spk} = \frac{\bar{\mathbf{z}}_{spk}}{\bar{\mathbf{z}}}, \quad (5)$$

where $\bar{\mathbf{z}}$ is the sample average of the VTR frequencies in the full training set and $\bar{\mathbf{z}}_{spk}$ is the sample average in the utterance from speaker spk. We now provide a brief analysis on this method. Using (3), we have

$$\bar{\mathbf{z}}_{spk} = \sum_s f_{spk,s} \bar{\mathbf{z}}_{spk,s} = \sum_s f_{spk,s} \boldsymbol{\beta}_{spk} \bar{\mathbf{z}}_s = \boldsymbol{\beta}_{spk} \sum_s f_{spk,s} \bar{\mathbf{z}}_s. \quad (6)$$

where $f_{spk,s} = \frac{n_{spk,s}}{\sum_s n_{spk,s}}$ is the relative frequency count, with $n_{spk,s}$ being the number of frames unit s is observed in the speaker-specific utterances, and $\bar{\mathbf{z}}_{spk,s}$ is the average VTR frequency of the phone s in speaker-specific utterances. Similarly, we have

$$\bar{\mathbf{z}} = \sum_s f_s \bar{\mathbf{z}}_s. \quad (7)$$

Then, (5) can be rewritten as

$$\bar{\boldsymbol{\beta}}_{spk} = \boldsymbol{\beta}_{spk} \frac{\sum_s f_{spk,s} \bar{\mathbf{z}}_s}{\sum_s f_s \bar{\mathbf{z}}_s}. \quad (8)$$

Eq. (8) indicates that the estimate of (5) would be accurate only when $\frac{\sum_s f_{spk,s} \bar{\mathbf{z}}_s}{\sum_s f_s \bar{\mathbf{z}}_s} = 1$, or $f_{spk,s} = f_s$; that is, when the frequency of HTM-unit s in the utterances of speaker spk is the same as that in the utterances of all training data. However, if these frequencies are different, then the estimate of (5) will be incorrect by a factor of $\frac{\sum_s f_{spk,s} \bar{\mathbf{z}}_s}{\sum_s f_s \bar{\mathbf{z}}_s}$. For example, in the extreme case when the utterance contains only unit of /aa/, then the scaling factor $\boldsymbol{\beta}_{spk}$ estimate becomes

$$\bar{\boldsymbol{\beta}}_{spk} = \boldsymbol{\beta}_{spk} \frac{\sum_s f_{spk,s} \bar{\mathbf{z}}_s}{\sum_s f_s \bar{\mathbf{z}}_s} = \boldsymbol{\beta}_{spk} \frac{\bar{\mathbf{z}}_{/aa/}}{\sum_s f_s \bar{\mathbf{z}}_s}. \quad (9)$$

On the other hand, the estimate from the utterance that contains only unit /iy/ would be

$$\bar{\boldsymbol{\beta}}_{spk} = \boldsymbol{\beta}_{spk} \frac{\sum_s f_{spk,s} \bar{\mathbf{z}}_s}{\sum_s f_s \bar{\mathbf{z}}_s} = \boldsymbol{\beta}_{spk} \frac{\bar{\mathbf{z}}_{/iy/}}{\sum_s f_s \bar{\mathbf{z}}_s}. \quad (10)$$

Since the average VTR frequencies of /aa/ and /iy/ are very different, the estimates (9) and (10) would differ vastly even though both utterances are generated by the same speaker.

3.2. Method two

The above problem is corrected by the second method described here, where the estimated scaling-factor is changed from (5) to

$$\bar{\boldsymbol{\beta}}_{spk} = \sum_s f_{spk,s} \frac{\bar{\mathbf{z}}_{spk,s}}{\bar{\mathbf{z}}_s}. \quad (11)$$

Now, substituting (3) into (11), we easily show that this estimate is not affected by the relative frequencies of the phonemes:

$$\bar{\boldsymbol{\beta}}_{spk} = \sum_s f_{spk,s} \boldsymbol{\beta}_{spk} = \boldsymbol{\beta}_{spk}. \quad (12)$$

4. Experimental Results

We have carefully analyzed the normalization results on the TIMIT database and observed that different speakers may have their VTR targets differ over as much as 40%. The range of the estimated target scaling factor $\bar{\boldsymbol{\beta}}_{spk}$ can be as low as 0.86 (female) and as high as 1.17 (male).

To further analyze the results, we have computed the root mean square (RMS) errors between the VTR trajectories estimated using the methods described in Section 3 and the results of a high-performance VTR tracker [3] over 192 TIMIT utterances. Table 1 summarizes the RMS error results when the scaling factors are estimated using the all units in these utterances. "Speaker independent" baseline VTR estimates (first row) were reported in [13], which incur very high RMS errors. The remaining two rows show RMS errors using the estimated targets by (2), where the scaling factor $\boldsymbol{\beta}_{spk}$ is estimated by (5) (Method 1) and by (11) (Method 2), respectively. Errors have been drastically reduced and the Method 2 has slightly lower errors, indicating a reasonably good data balance.

Table 2 summarizes the same comparative RMS error results as Table 1 except only the first five (instead of all) HTM-units in the utterances are used to estimate $\boldsymbol{\beta}_{spk}$. It is clear that target normalization makes a big difference in accurately predicting the VTR formants from the targets in both conditions especially for F2, F3 and F4. When the scaling factors are estimated using only the first 5 units in the utterances, the benefit of the second method becomes clear.

Table 1. *RMS errors for VTR trajectories using different ways of target estimation. Target scaling factors are estimated using all HTM-units in the utterances*

RMS Errors	F1 (Hz)	F2 (Hz)	F3 (Hz)	F4 (Hz)
Speaker independent	68	525	1010	1663
Speaker adjusted (1)	67	110	122	125
Speaker adjusted (2)	66	104	121	124

Table 2. *Same as Table 1 except target scaling factors are estimated using only the first five HTM-units in the utterances.*

RMS Error	F1 (Hz)	F2 (Hz)	F3 (Hz)	F4 (Hz)
Speaker independent	68	525	1010	1663
Speaker adjusted (1)	89	193	170	164
Speaker adjusted (2)	75	135	147	154

5. Complexity

In this section, we briefly discuss computational analysis on decoding speech using the HTM-based speech recognizer. We developed a time synchronous decoder for the HTM as detailed in [6]. In this decoder, the likelihood of each frame depends not only on the observation and HTM-unit identity associated with the current frame, but also on the unit identities associated with both previous and future D frames. That is, each hypothesis has to record all the HTM-unit identities for a 2D+1 frame-long window centered at the current frame, or $|\Sigma|^{2D+1}$ possibilities at each frame, where Σ is the size of the HTM-unit set. This gives the computation complexity of a naïve time synchronous decoder on the order of $O\left(T|\Sigma|^{2D+1}\right)$, where T is the total number of frames. We can restrict the search space by utilizing the lattices generated by an HMM and some carefully designed pruning strategies such as beam pruning and histogram pruning. After applying these technologies, the decoding time can be 100 times of the real time to achieve 74.68% phone accuracy for the TIMIT phone recognition task on the core test set [6].

6. Summary and Conclusion

HTM is one of two main types of structured statistical models developed in the past for automatic speech recognition. It uses non-recursive parameterization to characterize the long-span dependency of VTR as well as acoustic features in speech utterances. Like HMM, HTM is also a parametric model, but its structure is substantially more complex than HMM as well as other types of segment models [8]. Conventional wisdom says that more complex statistical models often would have a harder time to incorporate detailed knowledge into the algorithm development. As an example, in relatively simple models such as HMMs, speaker variation can be straightforwardly handled by pooling all data from many speakers in training. But for more complex models such as HTMs, such pooling would not work since some key parameter set (i.e., VTR targets) in the model are inherently speaker specific. While statistical distributions can be used to represent the randomness of the VTR targets due to speaker variation, this would significantly increase phonetic confusability. Special normalization techniques have been developed for HTM, as is the main focus of this paper.

7. References

[1] Deng, L., Yu, D., and Acero, A., "A quantitative model for formant dynamics and contextually assimilated reduction in fluent speech", Proc. Interspeech 2004.

[2] Deng, L., Li, X., Yu, D., and Acero, A., "A Hidden Trajectory Model with Bi-Directional Target-Filtering:" Proc. ICASSP 2005, pp 337-340.

[3] Deng, L., Acero, A., and Bazzi, I., "Tracking vocal tract resonances using a quantized nonlinear function embedded in a temporal constraint," IEEE Trans. Speech & Audio Processing, Vol. 14, No. 2, March 2006, pp. 425-434.

[4] Eide, E., and Gish, H., "A parametric approach to vocal tract length normalization," Proc. ICASSP 1996, pp. 346-348.

[5] Lee, K.F., and Hon, H.W., "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. Acoustics, Speech, and Signal Proc.*, Vol. 37, 1989, pp. 1641-1648.

[6] Li, X., Deng, L., Yu, D., Acero, A., "A Time-Synchronous Phonetic Decoder for a Long-Contextual-Span Hidden Trajectory Model", Proc. Interspeech 2006, pp. 609-612.

[7] McDonough, J., Byrne, W., and Luo, X., "Speaker normalization with all-pass transforms," Proc. ICSLP 1998, vol. 6, pp. 2307-2310.

[8] Ostendorf, M., Digalakis, V., and Rohlicek, J., "From HMMs to segment models: A unified view of stochastic modeling for speech recognition" IEEE Trans. Speech & Audio Proc., Vol. 4, 1996, pp. 360-378.

[9] Pye, D., and Woodland, P. C., "Experiments in speaker normalisation and adaptation for large vocabulary speech recognition", Proc. ICASSP 1997, pp. 1047-1050.

[10] Wegmann, S., McAllaster, D., Orloff, J., and Peskin, B., "Speaker normalization on conversational telephone speech," Proc. ICASSP 1996, pp. 339-341.

[11] Welling, L., and Haeb—Umbach, R., and Aubert, X., and Haberland, N., "A study on speaker normalization using vocal tract normalization and speaker adaptive training", Proc. ICASSP 1998, Vol. 2, pp. 797-800.

[12] Yu, D., Deng, L., and Acero, A., "Evaluation of a Long-contextual-span Hidden Trajectory Model and Phonetic Recognizer Using A* Lattice Search," Proc. Interspeech 2005, pp. 553-556.

[13] Yu, D., Deng, L., Acero, A., "Speaker-Adaptive Learning of Resonance Targets in a Hidden Trajectory Model of Speech Coarticulation", Computer Speech and Language, Vol. 27, 2007, pp. 72-87.

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