Learning Statistically Characterized Resonance Targets in a Hidden Trajectory Model of Speech Coarticulation and Reduction

Li Deng, Dong Yu, and Alex Acero

Microsoft Research
One Microsoft Way, Redmond, WA 98052, USA
{deng,dongyu,alexac}@microsoft.com

Abstract

We report our new development of a hidden trajectory model for co-articulated, time-varying patterns of speech. The model uses bi-directional filtering of vocal tract resonance targets to jointly represent contextual variation and phonetic reduction in speech acoustics. A novel maximum-likelihood-based learning algorithm is presented that accurately estimates the distributional parameters of the resonance targets. The results of the estimates are analyzed and shown to be consistent with all the relevant acoustic-phonetic facts and intuitions. Phonetic recognition experiments demonstrate that the model with more rigorous target training outperforms the most recent earlier version of the model, producing 17.5% fewer errors in N-best rescoring.

1. Introduction

In recent years, we have been developing a version of the statistical hidden trajectory model (HTM) where temporal filtering of vocal tract resonance targets is exploited as the basis for joint characterization of coarticulation and phonetic reduction in speech acoustics. As an extension of the stochastic segment models [1], our HTM embodies not only cross-frame correlation, but also cross-unit one, in the dynamic patterns of speech. A unique character of the HTM is the use of a highly compact set of context-independent parameters to capture the long-span context-dependent properties in acoustic features.

The scientific basis of the HTM was presented in [2], and a speech recognizer constructed using the HTM and its preliminary evaluation were described in [3]. A central concept in the HTM and the associated speech recognizer is the stochastic vocal tract resonance (VTR) target, where a (multivariate) probability distribution of the phone-dependent target vector is used to represent target variations across speakers and other factors (as well as co-variations among the target components). The parameters of the VTR target distribution require automatic training from data. In the work of [3], such training was empirical in that no clearly defined objective function is optimized. The estimates of the target means and variances were the sample statistics derived from the results of a previously developed VTR tracker [4]. In this paper, the training technique is improved by rigorous maximum likelihood (ML) estimation. Superior phonetic recognition results are obtained over the results reported in [3] based on more heuristic parameter estimation.

The organization of this paper is as follows. A complete and concise outline of the two-stage HTM is provided in Section 2. Key issues on implementing the model learning algorithm are discussed in Section 4, and experimental evaluation with detailed results is provided in Section 5.

2. Model Construction

2.1. Stochastic targets and their filtering

Stage-I of the HTM represents the time-varying pattern of stochastic hidden VTR vectors $\boldsymbol{z}_s$, whose smooth temporal movement is directed by statistically characterized target vectors $\boldsymbol{t}_s$ (subscript s denotes segmental phonetic unit). The generation of the VTR trajectories from the segmental targets is by a bi-directional finite impulse response (FIR) filtering:

$$\boldsymbol{z}_s(k) = h_{s(k)} * \boldsymbol{t}(k) = \sum_{\tau=k-D}^{k+D} c_\gamma \gamma_{s(\tau)}^{|k-\tau|} \boldsymbol{t}_{s(\tau)}, \qquad (1)$$

where c_γ is the normalization factor, which is needed to produce VTR target undershooting, instead of overshooting, for casually uttered speech. Parameter γ_s controls the *spatial* extent of coarticulation and is correlated with speaking effort. The length of the filter's impulse response $h_{s(k)}$, $2D+1$, determines the *temporal* extent of coarticulation.

The phone-dependent target vector $\boldsymbol{t}_s$ in (1) is a random vector — hence stochastic targets — whose distribution is assumed to be a (gender-dependent) multivariate Gaussian:

$$p(\boldsymbol{t}|s) = \mathcal{N}(\boldsymbol{t}; \boldsymbol{\mu}_{T_s}, \boldsymbol{\Sigma}_{T_s}). \qquad (2)$$

Then, due to the linearity between $\boldsymbol{z}$ and $\boldsymbol{t}$, the VTR vector $\boldsymbol{z}(k)$ (at each frame k) is also a Gaussian. Given a sampled target sequence $\boldsymbol{t}_{s(k)}$ from the distribution of (2), the model generates the random VTR trajectory $\boldsymbol{z}(k)$ with the Gaussian distribution:

$$p(\boldsymbol{z}(k)|s) = \mathcal{N}[\boldsymbol{z}(k); \boldsymbol{\mu}_{z(k)}, \boldsymbol{\Sigma}_{z(k)}] \qquad (3)$$

The mean vector of this Gaussian can be derived as

$$\boldsymbol{\mu}_{z(k)} = \sum_{\tau=k-D}^{k+D} c_\gamma \gamma_{s(\tau)}^{|k-\tau|} \boldsymbol{\mu}_{T_{s(\tau)}} = \mathbf{a}_k \cdot \boldsymbol{\mu}_T, \qquad (4)$$

and each f-th component of $\boldsymbol{\mu}_{z(k)}$ is

$$\mu_{z(k)}(f) = \sum_{l=1}^{L} a_k(l) \mu_T(l,f), \qquad (5)$$

where L is the total number of phone-like HTM units as indexed by l ($L = 58$ in our experiments), f=1,..., 8 for 4 VTR frequencies and 4 corresponding bandwidths.

The covariance matrix in (3) can be similarly derived to be

$$\boldsymbol{\Sigma}_{z(k)} = \sum_{\tau=k-D}^{k+D} c_\gamma^2 \gamma_{s(\tau)}^{2|k-\tau|} \boldsymbol{\Sigma}_{T_{s(\tau)}}.$$

Approximating the covariance matrix by a diagonal one for each phone unit l, we represent its diagonal elements as a vector:

$$\boldsymbol{\sigma}^2_{z(k)} = \boldsymbol{v}_k \cdot \boldsymbol{\sigma}^2_T. \tag{6}$$

where the target covariance matrix is also approximated as diagonal:

$$\boldsymbol{\Sigma}_T(l) \approx \begin{bmatrix} \sigma^2_T(l,1) & 0 & \cdots & 0 \\ 0 & \sigma^2_T(l,2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2_T(l,8) \end{bmatrix}$$

The f-th element of the vector in (6) is

$$\sigma^2_{z(k)}(f) = \sum_{l=1}^{L} v_k(l)\sigma^2_T(l,f). \tag{7}$$

In (5) and (6), $\mathbf{a}_k$ and $\boldsymbol{v}_k$ are frame (k)-dependent vectors. They are constructed for any given phone sequence and phone boundaries within the coarticulation range ($2D + 1$ frames) centered at frame k. (Any phone beyond the $2D + 1$ window contributes a zero value to the vectors' elements.) $a_k(l)$ is a function of $c(\gamma_{s(k)})\gamma_{s(\tau)}^{|k-\tau|}$, and $v_k(l)$ is a function of $c^2(\gamma_{s(k)})\gamma_{s(\tau)}^{2|k-\tau|}$. *They are both a function of the phones' identities and temporal orders in the utterance*, and are independent of the VTR dimension f.

2.2. Nonlinear cepstral prediction and its linearization

Stage-II of the HTM provides a probabilistic mapping or prediction from the stochastic VTR trajectory $\boldsymbol{z}(k)$ (output of model Stage-I) to the stochastic observation trajectory $\boldsymbol{o}(k)$. The observation takes the form of LPC cepstra in this paper. An analytical form of the nonlinear prediction function $\mathcal{F}[\boldsymbol{z}(k)]$ was presented in [3], and given this function, we represent the prediction residual cepstral vector as a multivariate Gaussian:

$$p(\boldsymbol{o}(k)|\boldsymbol{z}(k), s) = \mathcal{N}\Big[\boldsymbol{o}(k); \mathcal{F}[\boldsymbol{z}(k)] + \boldsymbol{\mu}_{r_{s(k)}}, \boldsymbol{\Sigma}_{r_{s(k)}}\Big]. \tag{8}$$

For computational tractability in marginalization over the VTR uncertainty (next section), it is desirable to linearize the nonlinear mean function of $\mathcal{F}[\boldsymbol{z}(k)]$ in (8). To accomplish this, we use the first-order Taylor series approximation to the nonlinear mean function:

$$\mathcal{F}[\boldsymbol{z}(k)] \approx \mathcal{F}[\boldsymbol{z}_0(k)] + \mathcal{F}'[\boldsymbol{z}_0(k)](\boldsymbol{z}(k) - \boldsymbol{z}_0(k)), \tag{9}$$

where the components of the Jacobian matrix can be computed in a closed form.

Substituting (9) into (8), we obtain the approximate conditional acoustic observation probability where the mean μ_{o_s} is expressed as a linear function of the VTR variable $\boldsymbol{z}$:

$$p(\boldsymbol{o}(k)|\boldsymbol{z}(k), s) \approx \mathcal{N}(\boldsymbol{o}(k); \boldsymbol{\mu}_{o_{s(k)}}, \boldsymbol{\Sigma}_{r_{s(k)}}), \tag{10}$$

where

$$\boldsymbol{\mu}_{o_{s(k)}} = \mathcal{F}'[\boldsymbol{z}_0(k)]\boldsymbol{z}(k) + \mathbf{b}_k,$$

with

$$\mathbf{b}_k = \mathcal{F}[\boldsymbol{z}_0(k)] - \mathcal{F}'[\boldsymbol{z}_0(k)]\boldsymbol{z}_0(k) + \boldsymbol{\mu}_{r_{s(k)}}.$$

2.3. Computing observation likelihood

Given the results above, we can now compute the likelihood of the acoustic observations (cepstra). This computation is essential because the likelihood provides a natural scoring mechanism comparing different linguistic hypotheses as needed in speech recognition. A closed-form computation is derived by marginalizing over hidden trajectories marginalization over the stochastic VTR vector $\boldsymbol{z}(k)$ as follows:

$$\begin{aligned} p(\boldsymbol{o}(k)|s) &= \int p[\boldsymbol{o}(k)|\boldsymbol{z}(k), s] p[\boldsymbol{z}(k)|s] d\boldsymbol{z} \\ \approx & \int \mathcal{N}[\boldsymbol{o}(k); \boldsymbol{\mu}_{o_{s(k)}}, \boldsymbol{\Sigma}_{r_{s(k)}}]\, \mathcal{N}[\boldsymbol{z}(k); \boldsymbol{\mu}_z(k), \boldsymbol{\Sigma}_z(k)] d\boldsymbol{z} \\ = & \ \mathcal{N}\Big\{\boldsymbol{o}(k); \bar{\boldsymbol{\mu}}_{o_{s(k)}}, \bar{\boldsymbol{\Sigma}}_{o_{s(k)}}\Big\} \end{aligned} \tag{11}$$

where the time-varying mean can be shown to be

$$\bar{\boldsymbol{\mu}}_{o_s}(k) = \mathcal{F}[\boldsymbol{z}_0(k)] + \mathcal{F}'[\boldsymbol{z}_0(k)][\mathbf{a}_k\boldsymbol{\mu}_T - \boldsymbol{z}_0(k)] + \boldsymbol{\mu}_{r_{s(k)}}$$

and the time-varying covariance matrix can be shown to be

$$\bar{\boldsymbol{\Sigma}}_{o_s}(k) = \boldsymbol{\Sigma}_{r_{s(k)}} + \mathcal{F}'[\boldsymbol{z}_0(k)]\boldsymbol{\Sigma}_z(k)(\mathcal{F}'[\boldsymbol{z}_0(k)])^{\mathrm{Tr}} \tag{12}$$

3. Model Learning

The goal of model learning presented in this section is to automatically estimate the model parameters, based on the cepstral observation data (no VTR data) in the training set, so as to maximize the observation likelihood in (11). The parameters of concern in this paper include all elements of the mean vectors and covariance matrices for both VTR targets and cepstral residuals.

3.1. Mean vectors in stochastic targets

To obtain a closed-form estimation solution, we assume diagonality of the prediction cepstral residual's covariance matrix $\boldsymbol{\Sigma}_{r_s}$. Denoting its j-th component by $\boldsymbol{\sigma}^2_r(j)$ ($j = 1, 2, ..., J$), we decompose the multivariate Gaussian of (11) element-by-element into

$$p(\boldsymbol{o}(k)|s(k)) = \prod_{j=1}^{J} \frac{1}{\sqrt{2\pi\sigma^2_{o_{s(k)}}(j)}} \exp\Big\{-\frac{(o_k(j) - \bar{\mu}_{o_{s(k)}}(j))^2}{2\sigma^2_{o_{s(k)}}(j)}\Big\}, \tag{13}$$

where $o_k(j)$ denotes the j-th component (i.e., j-th order) of the cepstral observation vector at frame k.

The log-likelihood function for a training data sequence ($k = 1, 2, ..., K$) relevant to the VTR mean vector μ_{T_s} becomes

$$\begin{aligned} P &= \sum_{k=1}^{K}\sum_{j=1}^{J}\Big\{-\frac{(o_k(j) - \bar{\mu}_{o_{s(k)}}(j))^2}{\sigma^2_{o_{s(k)}}(j)}\Big\} \qquad (14) \\ &= \sum_{k=1}^{K}\sum_{j=1}^{J}\Big\{\frac{[\sum_f \mathcal{F}'[z_0(k), j, f]\sum_l a_k(l)\mu_T(l,f) - d_k(j)]^2}{\sigma^2_{o_{s(k)}}(j)}\Big\} \end{aligned}$$

where l and f are indeces to phone and to VTR component, respectively, and

$$\begin{aligned} d_k(j) &= o_k(j) - b_k(j) = o_k(j) - \\ - & \ F[z_0(k), j] + \sum_f \mathcal{F}'[z_0(k), j, f] z_0(k, f) - \mu_{r_{s(k)}}(j). \end{aligned}$$

While the acoustic feature's distribution is Gaussian for both HTM and HMM given the state s, the key difference is that the mean and variance in HTM as in (14) are both time varying functions (hence trajectory model). These functions provide context dependency (and possible target undershooting) via the smoothing of targets across phonetic units in the utterance. This smoothing is explicitly represented in the weighted sum over all phones in the utterance (i.e., $\sum_l$) in (14).

Setting

$$\frac{\partial P}{\partial \mu_T(l_0, f_0)} = 0,$$

and grouping terms involving unknown $\mu_T(l, f)$ on the left and the remaining terms on the right, we obtain

$$\sum_f \sum_l A(l, f; l_0, f_0) \mu_T(l, f)$$

$$= \sum_k \Big\{ \sum_j \frac{\mathcal{F}'[z_0(k), j, f_0]}{\sigma^2_{o_{s(k)}}(j)} d_k(j) \Big\} a_k(l_0) \quad (15)$$

with $f_0 = 1, 2, ..., 8$ for each VTR dimension, and with $l_0 = 1, 2, ...58$ for each phone unit. In (15),

$$A(l, f; l_0, f_0) = \sum_{k,j} \frac{\mathcal{F}'[z_0(k), j, f] \mathcal{F}'[z_0(k), j, f_0]}{\sigma^2_{o_{s(k)}}(j)} a_k(l_0) a_k(l). \quad (16)$$

Eq. (15) is a 464×464 full-rank linear system of equations. Matrix inversion gives a ML estimate of the complete set of target distribution parameters: a 464-dimensional vector formed by concatenating all eight VTR components (four frequencies and four bandwidths)of the 58 units.

In implementing (15) for the ML solution to target mean vectors, we kept other model parameters constant. The estimation of the target and residual parameters was carried out in an iterative manner. Initialization of the parameters $\mu_T(l, f)$ was provided by the values in [5], which determines the initial Taylor series expansion points $z_0(k)$ in (15) and (16) for updating these target mean parameters.

3.2. Variances in statistical targets

Likewise, the log likelihood Eq.(11) for the cepstral observation sequences can be expressed as an explicit function of the stochastic targets' variances $\sigma^2_T(l_0, f_0)$. However, setting the gradient of this function to zero does not render a simple solution as for the target means above. We resort to the gradient ascent technique to optimize $\sigma^2_T(l_0, f_0)$. Details of the gradient computation are omitted here due to the space limit.

3.3. Cepstral residual means and variances

Estimation of the cepstral residual means and variances is identical to that outlined in [3] and omitted here. We note that these residual parameters provide an important mechanism for distinguishing speech sounds that belong to different manners of articulation. This is attributed to the fact that nonlinear cepstral prediction from VTRs has different accuracy for these different classes of sounds. Within the same manner class, the phonetic separation is largely accomplished by distinct VTR targets, which typically induce significantly different cepstral prediction values via a "amplification" mechanism provided by the Jacobian matrix $\mathcal{F}'[\boldsymbol{z}]$.

4. Learning-Algorithm Implementation

The major computation in the model training is the joint estimation of target mean vectors for all phones, with the rigorous solution derived in this paper and shown in Eq.(15). While the algorithm gives ML, instead of discriminative, learning results, the optimality is achieved only via joint processing of all training data, with simultaneous updating of parameters of all phones. This is in contrast to ML estimation for HMM that does not require joint processing of the data. The difference arises from the nature of the HTM that models long-span phonetic context.

The implementation of the target mean vector estimation solution provided in Eq.(15) requires the computation of a 464×464 matrix, whose elements are provided in Eq.(16), and the inverse of the matrix. The computational cost lies mainly in the need for accumulation over time frames in the training data for each element in the matrix; i.e., summation k in Eq.(16). We have devised two ways to reduce the computational cost. First, we re-organized the computation so that the computation of the j summation in Eq.(16) is cached for re-use whenever possible. This effectively reduced the computation by about 70% compared with brute-force implementation of Eq.(16). Second, we assumed block diagonality of the matrix, one block for each phone with dimensions of 8×8. (We have empirically observed that the elements outside the blocks are typically more than one order of magnitude smaller than those within the blocks.) This gave a total of 58 separate small matrices to accumulate and to invert. However, under most conditions, the convergence of such an approximate algorithm was not achieved and poor recognition results were obtained. The results presented in this section are therefore based on the large, full matrix solution of Eq.(15). Without the block diagonality approximation, the rigorous training over the full 4620 TIMIT training utterances takes 2.5 hours for each iteration when implemented in Matlab and run on a Pentium-IV machine. Convergence is reached typically within 4 iterations.

5. Experimental Evaluation

5.1. Experiment setup

We have carried out phonetic recognition experiments to evaluate the HTM with the new learning technique presented in this paper. The standard TIMIT database is used for the evaluation. A "flat" language model (i.e., bi-phone probabilities are set uniformly to one) is used in all experiments reported here. The standard TIMIT phone set with 48 labels is expanded to 58 (as described in [5]) in training the HTM parameters using the standard 4620 training utterances. Phonetic recognition errors are tabulated using the commonly adopted 39 labels after the folding. The results are reported on the standard core test set with a total of 192 utterances by 24 speakers.

The N-best rescoring paradigm is used to evaluate the HTM. For each of the core test utterances, a standard decision-tree-based triphone HMM, built in our lab with the bi-gram language model and MFCCs, is used to generate a large N-best list where $N = 1000$.

5.2. Results on phonetic recognition and analysis

As a baseline with the use of identical language models and of acoustic features, an HTK-implemented triphone HMM is built with a flat phone language model and LPC cepstral feature. Rescoring of the N-best list, with (N=1001) and without (N=1000) including reference hypotheses, gives the same accuracy of 64%. (Adding a bi-gram language model and replacing the LCP cepstra by Mel-cepstra improve the accuracy to 73%.)

The two HTM systems dramatically increase both phone and sentence recognition accuracies over the HMM system, as shown in Table 1. "Old-trn" refers to the system reported in [3] where the VTR target mean and variance parameters were trained based on sample statistics computed from 4620 sets of gender-dependent and speaker-adapted VTR target values derived by a VTR tracker of [4]. "New-trn" refers to the system trained with the method described in Section 3 of this paper in a gender-dependent manner and making no use of VTR tracking results. The new HTM training consistently outperforms the old training, especially for the N-best list with references included where an upper bound of performance is shown. For the $N = 1001$ list rescoring, the upper bound performance is improved by 17.5% in relative phone error rate reduction. Our analysis has identified a key factor accounting for the less significant performance improvement (1.9%) when no reference hypothesis is present in the N-best list shown in Table 1. That is, the high oracle error rate (18% error in our 1000-best list) has created many "holes" (incorrect phones) in the evaluated hypotheses. These "holes" are associated with incorrect VTR targets which undesirably reduce the acoustic scores of not only the incorrect phones where the "holes" are located but also the adjacent correct phones. Such an "error spreading" effect is a consequence of the long-contextual-span property of the HTM.

The simplest (but artificial) way to remove the "error spreading" effect is to include the reference hypothesis into the N-best list, as reported in Table 1. However, the number of competitive hypotheses (N=1000) in N-best rescoring paradigms is undesirably limited. This issue has been resolved by the work reported in the companion paper [6], which significantly extends the success of the HTM shown in Table 1 to billions of competitive hypotheses via efficient search over lattices.

Table 1: *Phonetic recognition performance comparison of the HTM with two training methods. Performance is measured by percent sentence and phone recognition accuracies (%) in the core test set of TIMIT. The same "Flat" language model is used for both types of HTM and for the HMM system. The acoustic features for all three systems are the same LPC cepstral vectors.*

Models	101-Best (with ref.)		1001-Best (with ref.)		1000-Best (no ref.)	
	sent	phn	sent	phn	sent	phn
HMM	0.0	64.0	0.0	64.0	0.0	64.0
HTM (old-trn)	83.3	95.6	78.1	94.3	0.5	73.0
HTM (new-trn)	85.9	96.4	81.8	95.3	0.5	73.5

5.3. Results on model learning

We show in this section typical results demonstrating the effectiveness of the training algorithm presented in Section 3. Selected VTR target mean (sub)-vectors before and after training are listed in Table 2. We note that whereas male and female gender-dependent VTR target mean vectors are initialized with the same values (pp. 364 in [5]), they become well separated after the training. And the estimated female's resonance frequencies are higher than the male's counterpart by an amount consistent with acoustic-phonetic intuition. Detailed analysis has been carried out to assess acoustic-phonetic reality of the training results and overwhelming consistency has been found.

Table 2: *VTR target frequency mean values before (B) and after (A) ML training (4 itr)*

	F1 (Hz)		F2 (Hz)		F3 (Hz)	
Phones	Male	Female	Male	Female	Male	Female
er (B)	490	490	1350	1350	1690	1690
er (A)	486	559	1382	1397	**1750**	**1832**
w (B)	350	350	770	770	2340	2340
w (A)	156	213	**835**	**874**	2275	2276
y (B)	360	360	2270	2270	2920	2920
y (A)	261	285	**2142**	**2240**	2897	3038
ax (B)	500	500	1500	1500	2500	2500
ax (A)	378	436	1559	1681	2482	2579

6. Discussion and Conclusions

The goal of the research presented in this paper is to develop a parsimonious speech model that captures the structure of underlying speech generation mechanisms and that performs better than the HMM for speech recognition, especially for free-style speech with strong coarticulation and phonetic reduction. We have recently extended the work of [3], which reported the initial development and evaluation of the HTM, in two significant ways. First, empirical target parameter learning used in [3] is improved by using rigorous ML learning based on the likelihood of the cepstral data, requiring no VTR trajectory data. The improved learning is presented in this paper, and is shown to have reduced phonetic recognition errors by 17.5% in N-best rescoring experiments. Second, the evaluation of the HTM is advanced from a relatively small-scale N-best rescoring (with N in the order of 1000) to a large-scale lattice search (equivalent to N in the order of billions in N-best lists). Details of this latter work is contained in the companion paper [6], which reports the effectiveness of the HTM and of the associated learning algorithm presented in this paper. Our future research involves further improving the quality of the current HTM as well as improving the efficiency of the HTM-specific search technique.

7. References

[1] M. Ostendorf, V. Digalakis, and J. Rohlicek. "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.

[2] L. Deng, D. Yu, and A. Acero, "A quantitative model for formant dynamics and contextually assimilated reduction in fluent speech", in Proc. ICSLP, pp 719-722, Jeju, Korea, 2004.

[3] L. Deng, X. Li, D. Yu, and A. Acero,"A hidden trajectory model with bi-directional target-filtering: Cascaded vs. integrated implementation for phonetic recognition", in Proc. ICASSP, pp 337-340, March 19-23, 2005, Philadelphia.

[4] 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. ICASSP, pp.557-560, Montreal, Canada, 2004.

[5] L. Deng and Doug O'Shaughnessy. SPEECH PROCESSING — A Dynamic and Optimization-Oriented Approach, Marcel Dekker Inc., New York, 2003.

[6] D. Yu, L. Deng, and A. Acero. "Evaluation of a long-contextual-span trajectory model and phonetic recognizer using A* lattice search", Interspeech 2005.