STATISTICAL MODELING OF CO-ARTICULATION IN CONTINUOUS SPEECH BASED ON DATA DRIVEN INTERPOLATION

Don X. Sun

Statistics and Information Analysis Research
Bell Laboratories, Lucent Technologies
Murray Hill, NJ 07974-0636, USA

ABSTRACT

Parsimonious modeling of the context dependency nature of speech due to co-articulation is very important for improving the performance of speech recognition systems. Numerous approaches have been proposed in the literature to address this problem. However, most of the methods are based on the idea of using context-dependent speech units, which inevitably increases the complexity of the model space. This paper presents a new approach of speech co-articulation modeling with complexity only comparable to context-independent models. In this model, the movement of a sequence of speech signals is characterized by a set of anchor points in the feature vector space that are corresponding to the target phonemic units. The transitions between the phonemic units due to co-articulation are modeled as interpolations between the target vectors. Two types of parameters are involved in the models: the intrinsic parameters in the models of target units and the auxiliary parameters specifying the transitional units. The auxiliary parameters are estimated "online" for a given sequence of speech feature vectors, hence it does not contribute to the complexity of the models. Unlike "triphone"-type context dependent models, the complexity of this approach is comparable to the context independent phoneme models, yet, some phonetic classification experiments showed that the new model can achieve the same performance as the more complex context dependent models.

1. INTRODUCTION

The context dependency nature of continuous speech due to co-articulation is a major problem for phoneme-based speech recognition. Although, the techniques of triphone/generalized triphone modeling ([10], [8], etc.) have been commonly used to deal with this problem, they often result a large number of models, which inevitably become difficult to train for limited amount of speech training data. Working towards parsimonious modeling of context dependency in continuous speech becomes extremely important for robust recognition. In [4], we have proposed a phonological feature based approach to reduce model space with parameter sharing among the states with similar articulatory features. Although the number of states is greatly reduced from the total number of all triphones, it is still much larger than the number of speech units in a context independent model.

In this paper, we propose a method of modeling the context dependency nature of speech without introducing more parameters than those in the context independent phonemic models. In this model, the movement of a sequence of speech signals is characterized by a set of anchor points in the feature vector space that are corresponding to the target phonemic units. The transitions between phonemic units due to co-articulation are modeled as interpolations between the target vectors. Two types of parameters are involved in the models: the intrinsic parameters in the models of target units and the auxiliary parameters specifying the transitional units. The auxiliary parameters are estimated "online" for a given sequence of speech feature vectors, hence it does not contribute to the complexity of the models. Unlike "triphone"-type context dependent models, the complexity of this approach is comparable to the context independent phoneme models.

This method shares the same idea of modeling dynamics movements of speech signal with many other methods developed in the recent years ([1], [5], [2], [7], [6], etc.). However, because the interpolation is not performed at the model levels, this approach does not increase model complexity for modeling the dynamics in speech.

2. MODEL FORMULATION

The basic idea underlying this model is to characterize a sequence of feature vectors as interpolations among a set of target units (or anchor points). The transitional portion of the speech is then modeled by smoothing spline based trajectories derived from the neighboring target units.

This model is motivated from the observation that although speech signal is highly dynamic, its movement tends to follow certain paths from one target to another corresponding to the underlying phonemic units. Figure 1 shows an example of three such paths for the phrase /all year/ in TIMIT. The paths are quite different from each other, however, the target positions at the upper-left and lower-right corners are rather consistent. This suggests that it is more important to model the target positions precisely than the paths of the intermediate positions that are merely results of the co-articulation phenomena.

Let $\mathbf{x}(t), t = 1, \cdots, T$ be a sequence of speech feature vectors calculated from a window of acoustic signals at time t, such as linear predictive coefficients, cepstrum coefficients etc. We usually take $\mathbf{x}(t)$ as a complete utterance. Let $(s_1, \cdots, s_k)$ be the underlying sequence of target phonemic units of the utterance $\mathbf{x}(t)$, and let $(\mathbf{y}(s_1), \cdots, \mathbf{y}(s_k))$ be the sequence of the target feature vectors corresponding to the phonemic units $(s_1, \cdots, s_k)$.

As in context-independent phonemic models, $\mathbf{y}(j)$ can either be modeled by a Gaussian distribution

$$\mathbf{y}(j) \sim N(\mu(j), \mathbf{\Sigma}(j))$$

or a mixture of Gaussian distributions

$$p(\mathbf{y}(j)|\mu(j), \mathbf{\Sigma}(j)) = \sum_{m=1}^{M} c(m) p_N(\mathbf{y}(j)|\mu(j,m), \mathbf{\Sigma}(j,m))$$

where $p_N(\cdot)$ is the density function of a single Gaussian distribution.

We model the sequence of speech signal $\mathbf{x}(t)$ by a smooth interpolating function plus a random noise part as follows:

$$\mathbf{x}(t) = \mathbf{f}(t; \mathbf{y}(s_1), \cdots, \mathbf{y}(s_k), t(s_1), \cdots, t(s_k)) + \epsilon(t), \quad (1)$$

where $(t(s_1), \cdots, t(s_k))$ is the sequence of knots for interpolation. The crucial component in this model is the interpolation between the target vectors $(\mathbf{y}(s_1), \cdots, \mathbf{y}(s_k))$, which can be modeled in many different ways. Amongth them, smoothing spline models interpolation is very desirable to use for two reasons: 1) the smoothness of the trajectories connecting various target units is automatically guaranteed; 2) the interpolation has explicit solution, which leads to very efficient computation [9, 12]. The general form of the cubic spline model is

$$\begin{aligned} \mathbf{f}(t|[t_{i-1}, t_i]) &= \frac{1}{6h_j} \Big[A_{i-1}(t_i - t)^3 + A_i(t - t_{i-1})^3 \\ &+ (6\mathbf{y}_{i-1} - A_{i-1}h_i^2)(t_i - t) \\ &+ (6\mathbf{y}_i - A_i h_i^2)(t - t_{i-1}) \Big] \end{aligned} \quad (2)$$

where A_i's are the second order derivatives and h_i's are the spacings between t_i's.

For simplicity, the spline model can also be replaced by a linear interpolating function:

$$\mathbf{f}(t|[t_{i-1}, t_i]) = \alpha \mathbf{y}_{i-1} + (1 - \alpha)\mathbf{y}_i$$

where $0 \leq \alpha \leq 1$.

3. ALGORITHM FOR PARAMETER ESTIMATION

The intrinsic model parameters to be estimated are $\mu(j)$ and $\mathbf{\Sigma}(j)$ for $j = 1, \cdots, J$, where j represents j-th phonemic unit, and J is the total number of such units in a given speech recognition task. For each given utterance $\mathbf{x}(t)$, we also need to estimate the auxiliary parameters $(\mathbf{y}(s_1), \cdots, \mathbf{y}(s_k), t_1, \cdots, t_k)$. The problem of estimating the model parameters can be solved by maximizing the following objective function:

$$\begin{aligned} & L(\mathbf{x}|\mathbf{y}, \mu, \mathbf{\Sigma}) \\ = & \sum_{j=1}^{k} \log(p(\mathbf{y}(s_j)|\mu(s_j), \mathbf{\Sigma}(s_j))) \\ & - \sum_{t=1}^{T} ||\mathbf{x}(t) - \mathbf{f}(t; \mathbf{y}(s_1), \cdots, \mathbf{y}(s_k), t(s_1), \cdots, t(s_k)||^2 \\ = & A(\mathbf{y}, \mu, \mathbf{\Sigma}) - B(\mathbf{x}, \mathbf{y}) \end{aligned} \quad (3)$$

where the second term only involves the auxiliary parameters $(\mathbf{y}(s_1), \cdots, \mathbf{y}(s_k), t(s_1), \cdots, t(s_k))$.

3.1. Training

For a given set of n utterances $\mathbf{x}^i(t)$, $i = 1, \cdots, n$ with known phonemic transcriptions $(s_1^i, \cdots, s_{k(i)}^i)$ for each i, the objective function becomes

$$L(\mathbf{x}^1, \cdots, \mathbf{x}^n|\mathbf{y}^1, \cdots, \mathbf{y}^n, \mu, \mathbf{\Sigma}) = \sum_{i=1}^{n} \left(A(\mathbf{y}^i, \mu, \mathbf{\Sigma}) - B(\mathbf{x}^i, \mathbf{y}^i) \right). \quad (4)$$

The maximization of this objective function can be achieved by the following alternating iterative procedure:

1. Obtain initial estimates $(\mu, \mathbf{\Sigma})$ from the standard phonemic hidden Markov models.
2. For given $(\mu, \mathbf{\Sigma})$, estimate the optimal knot sequence $(t_1^i, \cdots, t_{k(i)}^i)$ for $(s_1^i, \cdots, s_{k(i)}^i)$ and the corresponding target vectors $(\mathbf{y}^i(s_1^i), \cdots, \mathbf{y}^i(s_{k(i)}^i))$ separately for each $\mathbf{x}^i(t)$.
   The estimation is achieved by an iterative procedure, where in each step, only one knot is optimized for $B(\mathbf{x}^i, \mathbf{y}^i)$ with all other knots fixed. Good initial knots can be obtained from the middle points of each phoneme segment or, even better, the change points in the feature space.
3. For fixed $(\mathbf{y}^i(s_1^i), \cdots, \mathbf{y}^i(s_{k(i)}^i))$, the estimation of $\mu(j)$ and $\mathbf{\Sigma}(j)$ does not involve the second term $B(\mathbf{x}^i, \mathbf{y}^i)$ in (4). $\mu(j)$ and $\mathbf{\Sigma}(j)$ can be estimated by maximizing the j-th component of the log-likelihood function in $\sum_{i=1}^{n} A(\mathbf{y}^i, \mu, \mathbf{\Sigma})$.

Figure 2 shows the estimated target vectors as well as the interpolation between the target points for the examples in Figure 1.

3.2. Recognition

For a given utterance $\mathbf{x}(t)$, $t = 1, \cdots, T$, we want to estimate the phonetic transcription $(s_1, \cdots, s_k)$ as well as the number of phonemic units k. The steps of decoding the phonetic sequence are as follows:

1. Set initial value of $k = 1$.
2. Find a sequence of knots $(t_1, \cdots, t_k)$ with their corresponding target vectors $(\mathbf{y}(t_1), \cdots, \mathbf{y}(t_k))$ such that $\sum_{t=1}^{T} ||\mathbf{x}(t) - \mathbf{f}(t; \mathbf{y}(t_1), \cdots, \mathbf{y}(t_k), (t_1, \cdots, t_k))||^2$ is minimized. This is essentially the same as step 2 for training.
3. For each $\mathbf{y}(t_i)$, find
   $$s_i = \text{argmax}\ {}_{j=1}^{J} p(\mathbf{y}(t_i)|\mu(j), \mathbf{\Sigma}(j)).$$
4. Set $k \leftarrow k + 1$, and repeat the procedure until the phonetic sequence $\mathbf{s} = (s_1, \cdots, s_k)$ remains the same (assuming that consecutive duplicated symbols in $\mathbf{s}$ are removed).

4. MAIN RESULTS

The TIMIT database is chosen for the evaluation experiments. As a first step, we only performed phoneme classification experiments in comparison with some other models.

The training subset consists of 100 speakers and the testing subset consists of 40 speakers. The results of the experiments are phonetic classification of the 61 TIMIT quasi-phonemic labels folded into 39 classes. Seven Mel-frequency cepstral coefficients (MFCC) and their differences are used as speech feature vectors.

The proposed method is compared with both context independent models and context dependent models described as follows:

CI-N3M5 Context independent phoneme models with 3 HMM states for each phoneme and 5 Gaussian mixture components per state.
(Total number of Gaussian kernels: 61x3x5= 915.)

CI-N3M100 Context independent phoneme models with 3 HMM states per phoneme and 100 Gaussian mixture components per state. The reason for including this model is to compare with the next context dependent model with similar number of parameters.
(Total number of Gaussian kernels: 61x3x100= 18300.)

CD-N3M5 The context dependent units are chosen to be the generalized triphones as described in [11].
(Total number of Gaussian kernels: 39x(15+15+1)x3x5= 18135.)

DDI Date driven interpolation models with one state per phoneme and 15 mixture components per state model. (Total number of Gaussian kernels: 61x15= 915.)

Model Type	Accuracy Rate
CI-N3M5	62.8%
CI-N3M100	64.6%
CD-N3M5	73.4%
DDI (Proposed method)	72.5%

Table 1. TIMIT phoneme classification experiment

From this experiment, it is quite clear that context dependent models outperform context independent models by a significant amount. By merely increasing the model complexity in a context independent model does not help too much. Interestingly, with the proposed method, we are able to achieve the same performance as the generalized triphones with much smaller number of model parameters.

5. DISCUSSIONS

In this study, we have proposed a new method of modeling speech co-articulation without the use of context dependent speech units. This approach has a few advantages: First, the distributions of the target phonemic units are much tighter than those estimated from the hidden Markov models with all the frames included for parameter estimation. Second, the memory requirement for this new model is much less demanding comparing with the models involving context-dependent speech units. Third, the recognition step does not require dynamic programming based search procedure, and it can run for any given length of a partial utterance. Therefore it can be very suitable for real-time speech recognition tasks.

REFERENCES

[1] L. Deng. A generalized hidden markov model with state-conditioned trend function of time for the speech signal. *Signal Processing*, 27(1):65–78, 1992.

[2] L. Deng, M. Aksmanovic, D. X. Sun, and C. F. J. Wu. Speech recognition using hidden markov models with polynomial regression functions as nonstationary states. *IEEE Transactions on Speech and Audio Processing*, 2(4):507–520, 1994.

[3] L. Deng, P. Kenny, and P. Mermelstein. Modeling acoustic transitions in speech by state-interpolation hidden markov models. *IEEE Trans. Signal Processing*, 40(2):265–271, 1992.

[4] L. Deng and D. X. Sun. A statistical approach for automatic speech recognition using the atomic units constructed from overlapping articulatory features. *Journal of the Acoustical Society of America*, 95(5):2702–2719, 1994.

[5] O. Ghitza and M. Sondhi. Hidden markov models with templates as nonstationary states: an application to speech recognition. *Computer Speech and Language*, 7(2):101–119, 1993.

[6] W. Goldenthal. *Statistical Trajectory Models for Phonetic Recognition.* PhD thesis, Massachusetts Institute of Technology, 1994.

[7] Y. Gong and J. P. Haton. Stochastic trajectory modeling for speech recognition. In *Proceedings of Int. Conf. on Acoustics, Speech and Signal Processing, Adelaide, Australia*, pages 57–60, 1994.

[8] M. Hwang, X. Huang, and F. Alleva. Predicting unseen triphones with senones. In *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, pages 311–315, 1993.

[9] P. Lancaster and K. Salkauskas. *Curve and Surface Fitting: An Introduction.* Academic Press, 1986.

[10] K. Lee and H. Hon. Speaker-independent phone recognition using hidden markov models. *IEEE Trans. Signal Processing*, 38(4):599–609, 1990.

[11] C. Rathinavelu and L. Deng. Use of generalized dynamic feature parameters for speech recognition: maximum likelihood and minimum classification error approaches. In *Proceedings of Int. Conf. on Acoustics, Speech and Signal Processing*, pages 373–376, 1995.

[12] D. X. Sun. Robust estimation of spectral center-of-gravity trajectories using mixture spline models. In *Proceedings of the 4th European Conference on Speech Communication and Technology, Madrid, Spain*, pages 749–752, 1995.

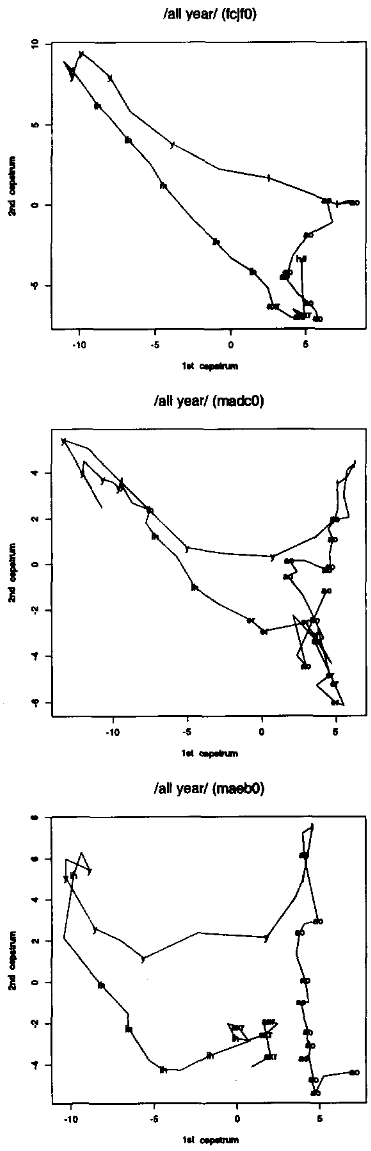

Figure 1. First two cepstrum coefficients of three different realizations of the phrase /all year/ in TIMIT

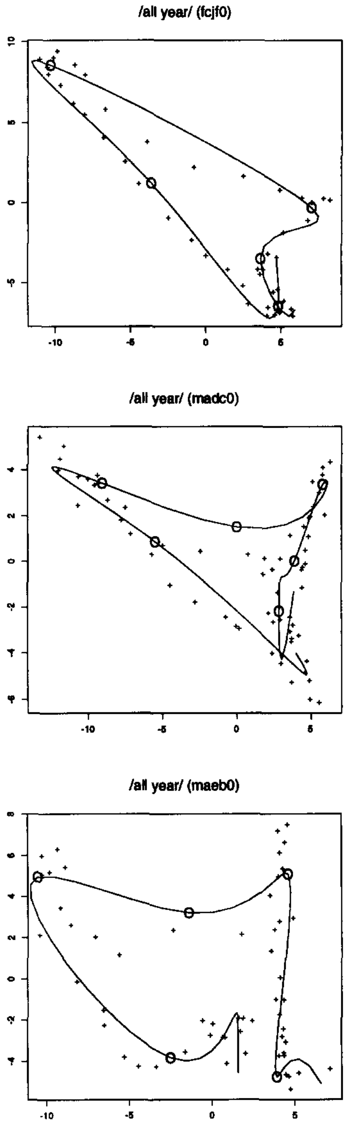

Figure 2. Interpolation between target vectors by smoothing spline. The circles represent the estimated target vectors.