Unified Frame and Segment Based Models for Automatic Speech Recognition

Hsiao-Wuen Hon and Kuansan Wang

Speech Technology Group, Microsoft Research
Redmond, Washington 98052, USA

ABSTRACT

In this paper, we propose an analytically tractable framework that integrates the frame and segment based acoustic modeling techniques. We combine the two approaches by jointly modeling their respective hidden Markov processes. Since the joint process is based on the same mathematical framework, conventional search and training techniques, such as Viterbi and EM algorithms, can be directly applied. It also allows the score from either model to contribute to the training and decoding of the other, reaching a jointly optimal decision. We conducted two series of experiments to verify our hypotheses. In the phone-pair classification experiments, our segment models show a 24% error reduction over state-of-the-art HMM-based system. The superior quality of segment models contributes to an 8.2% reduction in word error rates for the unified system on the WSJ dictation task.

1. INTRODUCTION

In laying out a pattern recognition framework for variable-length patterns such as speech, one usually assumes the pattern is generated by an unobservable Markov chain of basic units, each of which captures the local salient features of the signal and can be modeled by a probabilistic distribution with fixed parameters. The frame-based approaches further assume that, within each unit, the acoustic frames are independent and stationary. While these assumptions lead to tractable and efficient implementations, they are inconsistent with the properties of speech [4,6]. Segment based methods [1-5], on the other hand, do not employ the piecewise stationary and conditional independence assumptions. However, these systems are usually more expensive and, although they have produced encouraging performance for small vocabulary or isolated recognition tasks, their effectiveness on large vocabulary continuous speech recognition (LVCSR) remains an open issue.

The strengths of the two approaches are complementary and can be integrated seamlessly. HMMs are very effective in modeling the subtle details of speech signals by using one state for each quasi-stationary region. However, the transitions between quasi-stationary regions are largely neglected by HMMs. In contrast, segment models (SM) are powerful in modeling the transitions and longer-range speech dynamics, but often require compromises to reduce the complexity of implementation. In this paper, we report our unified framework of combining HMMs and SMs.

This paper is organized as follows. In Section 2 we present the general framework. In Section 3 we then discuss our SM approach based on parametric mixture trajectory models. In Section 4 we report a classification experiment aimed at verifying the effectiveness of our SM. In Section5, we present the recognition results on a LVCSR task using the combination framework. Finally, major findings are summarized and future works are discussed in Section 6.

2. MATHMEMATICAL FRAMEWORK

In the statistical approach to automatic speech recognition, the mathematical optimal solution dictates that the recognizer follow the maximum *a posteriori* (MAP) decision rule

$$\hat{\mathbf{W}} = \arg\max\, p(\mathbf{W} \mid \mathbf{O}) = \arg\max\, p(\mathbf{O} \mid \mathbf{W})\, p(\mathbf{W}) \qquad (1)$$

where $\mathbf{W}$ is a word string hypothesis for a given acoustic observation $\mathbf{O}$. When deriving the acoustic model score $p(\mathbf{O} \mid \mathbf{W})$, a hidden process $\mathbf{q}$ is usually introduced as

$$p(\mathbf{O} \mid \mathbf{W}) = \sum_{\mathbf{q}} p(\mathbf{O} \mid \mathbf{q})\, p(\mathbf{q} \mid \mathbf{W}) \qquad (2)$$

in which it is assumed the hidden process can fully account for the conditional probability of the acoustic signal.

In the frame-based HMM approach, the hidden process is a path through a Markov chain, and the probabilistic distribution of an acoustic frame $\mathbf{o}_t$ is determined by the Markov state in which the hidden process (denoted $\mathbf{q}^h$ herein) is at time *t*. It is common to assume that (i) the underlying Markov chain is of order one, (ii) the probability distribution for the acoustic observation remains the same over time within the same Markov state, and that (iii) the acoustic observations are statistically independent. These assumptions lead to

$$p(\mathbf{O} \mid \mathbf{q}^h)\, p(\mathbf{q}^h \mid \mathbf{W}) = \prod_t p(\mathbf{o}_t \mid q_t^h)\, p(q_t^h \mid \mathbf{W}) \qquad (3)$$

a formulation that is tractable and has many desirable properties for implementation. As mentioned before, these assumptions have drawbacks that may be alleviated by including segment models. If we denote the hidden process of a segment model as $\mathbf{q}^s$, the combination can be achieved through

$$\begin{aligned} p(\mathbf{O} \mid \mathbf{W}) &= \sum_{\mathbf{q}^h} \sum_{\mathbf{q}^s} p(\mathbf{O}, \mathbf{q}^h, \mathbf{q}^s \mid \mathbf{W}) \\ &= \sum_{\mathbf{q}^h} \sum_{\mathbf{q}^s} p(\mathbf{O} \mid \mathbf{q}^h, \mathbf{q}^s)\, p(\mathbf{q}^s \mid \mathbf{q}^h)\, p(\mathbf{q}^h \mid \mathbf{W}) \end{aligned} \qquad (4)$$

In essence, we combine the frame and segment based approaches by joining their hidden processes. By treating the combination as a hidden data problem, one can apply the decoding and iterative training techniques commonly used in HMMs and SMs to the combination framework as well. By tightly coupling the two hidden processes, the HMM re-estimation and decoding may also benefit from its segmental counterpart, and vice versa.

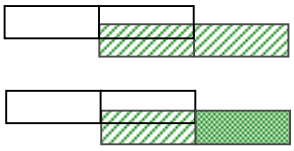
In this work, we let the conditional probability of the acoustic signal be

$$p(\mathbf{O} \mid \mathbf{q}^s, \mathbf{q}^h) = p(\mathbf{O} \mid \mathbf{q}^s)^a \, p(\mathbf{O} \mid \mathbf{q}^h) \quad (5)$$

where a is a constant weight. Following the same rationale in [9], segment recognition can be performed by detecting and decoding a sequence of salient events in the acoustic stream. These events mark the boundaries of speech segments, which are assumed to be statistically independent. In other words,

$$p(\mathbf{O} \mid \mathbf{q}^s) = \prod_i p(\mathbf{Y}_i \mid q_i^s) \quad (6)$$

where $\mathbf{Y}_i$ denotes the i^{th} segment. Because of the independent segment assumption, we believe the segment units should not be shorter than a phone.



(a) phone-pair SM

(b) two phone (monophone or gen. triphone) SM's

Figure 1 Overlapped evaluation where (a) a phone-pair SM is used, or (b) back off to two phones when the phone-pair (e_{i-1}, e_i) SM does not exist.

In this paper, three types of units are used: monophone, generalized triphone, and phone-pair (which is composed of two adjacent phones). The probabilistic distribution for each segment employs the trajectory modeling methods described in Section 3. Suppose e_i and t_i denote the phone and the starting time of the i^{th} segment. For the monophone or generalized triphone segment, a trajectory model is trained for the acoustic segments between t_i and t_{i+1}, i.e.,

$$p(\mathbf{Y}_i \mid q_i^s) = p(\mathbf{o}_{t_i}^{t_{i+1}-1} \mid e_i) \quad (7)$$

For the phone-pair segment, we have a trajectory model straddling across the phone boundary, as shown in Figure 1 (a):

$$p(\mathbf{Y}_i \mid q_i^s) = p(\mathbf{o}_{t_{i-1}}^{t_{i+1}-1} \mid e_{i-1}, e_i) \quad (8)$$

In the case for which a phone-pair (e_{i-1}, e_i) does not have enough occurrences in the training data, we approximate the phone-pair model with its monophone or generalized triphone back off, as shown in Figure 1 (b), namely,

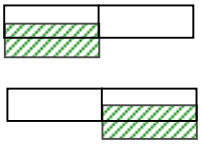
$$p(\mathbf{Y}_i \mid q_i^s) = p(\mathbf{o}_{t_{i-1}}^{t_i-1} \mid e_{i-1}) \, p(\mathbf{o}_{t_i}^{t_{i+1}-1} \mid e_i) \quad (9)$$

As evident in Figure 1, using phone-pair as the segment unit results in an "*overlapped evaluation*," which means the contribution for each acoustic frame is counted twice towards the overall score. For either the beginning or ending phone segment in a sentence, an extra phone based SM evaluation is performed as shown in Figure 2, to assure completely overlapped evaluation. Overlapped evaluation also free us from additional normalization for Eq. (5) We shall elaborate the significance of overlapped evaluation in Section 6.

Finally, it is also assumed here that the phone sequence and phone boundaries hypothesized by the HMM and segment model agree with each other. Based on the independent segment assumption, this leads to a segment duration model

$$p(\mathbf{q}^s \mid \mathbf{q}^h) = \prod_i p(t_i, t_{i+1} - 1 \mid e_i), \quad (10)$$

which, as in [9], is modeled by an ergodic Poisson mixture distribution. This Unified framework enables both frame and segment based models to equally contribute to *jointly* optimal segmentations, which should lead to more efficient pruning during search. The inclusion of the SM should not require massive changes in the decoder's design because the SM scores can be handled in the same manner as the language model scores. Finally, Eq. (4) can be used to estimate the model parameters for both HMM and SM for joint optimalization.



(a) the beginning phone

(b) the ending phone

Figure 2. Overlapped evaluation for the beginning and ending phone segments in a sentence.

3. MIXTURE TRAJECTORY MODELS

In this section, we present our SM approach, which is based on parametric *trajectory segment models* [2,4]. Our SM approximates the D-dimensional acoustic observation vector $\mathbf{Y} = (\mathbf{y}_0, \mathbf{y}_1, \cdots, \mathbf{y}_{T-1})$ by a polynomial function. Specifically, the observation vector $\mathbf{y}_t$ can be represented as

$$\begin{aligned} \mathbf{y}_t &= \mathbf{C} \times F_t + \mathbf{e}_t(\Sigma) \\ &= \begin{pmatrix} c_1^0 & c_1^1 & \cdots & c_1^N \\ c_2^0 & c_2^1 & \cdots & c_2^N \\ \vdots & \vdots & \vdots & \vdots \\ c_D^0 & c_D^1 & c_D^2 & c_D^N \end{pmatrix} \begin{pmatrix} f_0(t) \\ f_1(t) \\ \vdots \\ f_N(t) \end{pmatrix} + \mathbf{e}_t(\Sigma) \end{aligned} \quad (11)$$

where the matrix $\mathbf{C}$ is the trajectory parameter matrix, F_t is the family of N^{th} order polynomial functions, and $\mathbf{e}_t(\Sigma)$ is the residual fitting error. Eq. (11) can be regarded as modeling the time-varing mean in the output distribution for a HMM state. To simplify computation, the distribution of the residual error is often assumed to be an independent and identically distributed random process with a normal distribution $N(0, \Sigma)$. To accomdate diverse durations for the same segment, we follow the same approach as in [2] by using the relative linear time sampling of the fixed trajectory.

Each segment M is characterized by a trajectory parameter matrix $\mathbf{C}_m$ and covariance matrix Σ_m. The likelihood for each frame can be specified as

$$P(\mathbf{y}_t \mid \mathbf{C}_m, \Sigma_m) = \frac{\exp\left\{-tr\left[(\mathbf{y}_t - \mathbf{C}_m F_t)\Sigma_m^{-1}(\mathbf{y}_t - \mathbf{C}_m F_t)^t\right]/2\right\}}{(2\pi)^{D/2} \mid \Sigma_m \mid^{1/2}} \quad (12)$$

If we let $\mathbf{F} = (F_0, F_1, \cdots, F_{T-1})^t$, then the likelihood for the whole acoustic observation vector $\mathbf{Y}$ can be expressed as

$$P(\mathbf{Y}\mid\mathbf{C}_m,\Sigma_m)=\frac{\exp\left\{-tr\left[(\mathbf{Y}-\mathbf{C}_m\mathbf{F})\Sigma_m^{-1}(\mathbf{Y}-\mathbf{C}_m\mathbf{F})^t\right]/2\right\}}{(2\pi)^{DT/2}\mid\Sigma_m\mid^{T/2}} \quad (13)$$

Multiple mixture components can also be applied to SM. Suppose segment *M* is modeled by *K* trajectory mixtures. The likelihood for the acoustic observation vector $\mathbf{Y}$ in Eq. (13) becomes

$$\sum_{k=1}^{K} w_k\, p_k(\mathbf{Y}\mid\mathbf{C}_k,\Sigma_k). \quad (14)$$

Assuming an a sample of L observation vectors $\mathbf{Y}_1,\mathbf{Y}_2,\cdots,\mathbf{Y}_L$ with corresponding duration $T_1,T_2,\cdots,T_L$ are accounted for by the segment model *M*. The maximum likelihood re-estimation formulae using the EM algorithm can be derived as follows:

$$\gamma_k^i=\frac{w_k\,p_k(\mathbf{Y}_i\mid\mathbf{C}_k,\Sigma_k)}{P(\mathbf{Y}_i\mid\Phi_m)} \quad (15)$$

$$\hat{w}_k=\frac{1}{L}\sum_{i=1}^{L}\frac{w_k\,p_k(\mathbf{Y}_i\mid\mathbf{C}_k,\Sigma_k)}{P(\mathbf{Y}_i\mid\Phi_{\mathrm{m}})} \quad (16)$$

$$\hat{\mathbf{C}}_k=\left[\sum_{i=1}^{L}\gamma_k^i\mathbf{Y}_i\mathbf{F}_{T_i}^t\right]\left[\sum_{i=1}^{L}\gamma_k^i\mathbf{F}_{T_i}\mathbf{F}_{T_i}^t\right]^{-1} \quad (17)$$

$$\hat{\Sigma}_k=\sum_{i=1}^{L}\gamma_k^i(\mathbf{Y}_i-\mathbf{C}_k\mathbf{F}_{T_i})(\mathbf{Y}_i-\mathbf{C}_k\mathbf{F}_{T_i})^t\Big/\sum_{i=1}^{L}\gamma_k^i T_i \quad (18)$$

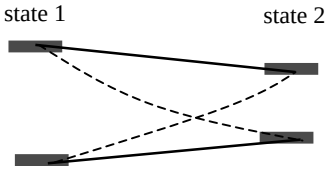


Figure 3. Diagram illustrates that HMM's output observation can hop between two unexpected quasi-stationary states.

While our formulations (Eq. (11) – (18)) are similar to [6], they are more efficient to implement because no individual polynormial fitting is required for each training segment. Segment mixtures can be considered as a complement to Gaussian mixtures in HMMs [7]. Figure 3 shows a digram where each of the two adjacent Markov states has two mixture components. The solid lines denote the valid trajectories actually observed in the data. However, HMM will inadvently allow two *phantom* trajectories, as shown in Figure 3 with dashed lines. HMM's independent observation assumption places no constraint on the mixture component transitions across the states. It is likely that such phantom trajectories are overgenerated and do not exist in real data. The segment mixture can be used to model the two valid trajectories exclusively when the two states are modeled within the same segment.

4. PHONE-PAIR CLASSIFICATION

To assess the usefulness of our segment models described in the previous section, a phone-pair classification experiment is first conducted. The continuous female Wall Street Journal (WSJ) training database is automatically segmented into a phone-pair database with the help of the Whisper system [8]. To ensure that there is enough training data, we only test the 500 most frequent phone-pairs in the database. These 500 most frequent phone-pairs comprise about two thirds of the entire female WSJ training data. There are between 300 to 33458 occurrences of these 500 phone-pairs.

We use Whisper as the baseline frame-based system, where each phone-pair is modeled by a 6-state HMM. Since there is enough training data for each phone-pair HMM, each state uses 20 Gaussian mixtures with no sharing. The acoustic feature used for the baseline is a 33-dimension vector consisting of 12 MFCC, 12 delta-MFCC, 6 delta-delta-MFCC, power, delta-power, and delta-delta-power. The baseline has a 31.5% classification error. For the trajectory models described in Section 3, we use only 12 MFCC and power as the feature vector to see how well the trajectory can model the dynamic features. The classification result is disappointing, as indicated in Table 1.

System	Order	Mix.	Co-Var. Matrix	Error Rate
HMM				31.5%
SM1	3	5	F	37.2%
SM2	5	7	F	33.9%
SM3	5	10	F	34.1%
SM4	5	10	D	34.2%

Table 1 Phone-pair classification results for segment models with MFCC+power only. The second and the third columns show the order of the polynomial and the number of trajectory mixtures used in our SM. 'F' and 'D' in the fourth column indicate whether a full or diagonal covariance matrix is used, respectively.

Next, we add the delta features (but not the delta-delta ones). The classification results are much better than the HMM baseline and the best configuration shows a 24% error reduction over the baseline system as seen in Table 2. This indicates that the fixed-time speech dynamics like delta features are essential and are not well modeled by the linear time sampling in our segment models.

System	Order	Mix.	Co-Var. Matrix	Error Rate	Error Reduction
HMM				31.5%	N/A
SM1	3	7	F	24.9%	21%
SM3	5	7	F	23.9%	24%
SM4	5	10	F	24.2%	23%
SM5	5	10	D	26.9%	15%
SM6	5	15	D	25.7%	18%

Table 2 Phone-pair classification results for segment models with additional dynamic features.

5. LVCSR EXPERIMENTS

The phone-pair classification results depict a very promising future for the trajectory model, considering the SM uses a smaller set of features (no delta-delta features) and fewer model parameters than HMM. For our preliminary LVCSR experiments conducted on the same female WSJ database, we use the WSJ 5K

dictation task and the ARPA Lincoln Lab 5K bigram language model. To have a larger test set, we combine November 92 5K-word dictation evaluation test set and 94 s0 development set. The baseline system uses 6000 senones, in which each senone is modeled by 20 Gaussian mixtures.

Our preliminary experiments use an N-best rescoring technique as in [5]. The top 100-best sentence hypotheses are generated by the HMM, and these hypotheses are then re-scored based on the combination of HMM and segment model scores using Eq (5). For SMs, we generate the 850 most frequent phone-pair models and all the monophone models for back off. For better back-off models, we create 2000 generalized triphone SMs using a triphone decision tree. The results are summarized in Table 3.

	ERROR	ERROR REDUCTION
Baseline HMM	7.19%	N/A
+phone-pair SM +monophone	6.97%	3.3%
+triphone SM	6.78%	5.7%
+phone-pair SM +triphone SM	6.60%	8.2%

Table 3 WSJ dictation results using various segment models. The 3rd row shows the results using only general triphone segments in conjunction with HMM, where the 2nd and 4th rows use phone-pairs with monophone and general triphone back off, respectively.

In the test set, there are about 25% of the phone-pair units requiring phone back off. By analyzing the results, it is clear that context-dependency is important to SM as well. Naturally, we should be able to improve the results further by using context dependent phone-pair models. Moreover, variance smoothing seems to be critical. Current experiments used the simplest grand variance. We are expecting to leverage all the advanced smoothing techniques, such as MAP [10], to make the SM more robust. Finally, our preliminary results have not taken full advantage of the unified framework as outlined in Section 2. Ongoing work is being conducted to decode the jointly optimal results and carry out iterative training accordingly.

6. CONCLUSIONS

In this article, we present a unified frame and segment based model for ASR. The merits of the segment models are first verified in phone-pair classification experiments. Our results show 24% improvement in the quality of the acoustic model, which leads to an 8.2% improvement in WSJ dictation task.

There are three unique and significant findings in this work, the first of which is the use of long and supra-phonetic segments. Our results strongly support what many speech scientists have advocated that long-range acoustic evolution patterns in the time-frequency domain are critical for speech recognition. However, the realization of segment models often involves a considerable amount of complexity, and reasonable compromises made for implementation often wipe out the potential gains they stand to offer. We hope the unified framework proposed here will manifest itself as a reasonable yet effective compromise.

Secondly, our results show that only a handful target trajectories are needed for speaker independent recognition. Once again, we view this outcome as supporting evidence for the invariant feature hypothesis advocated in phonetic target theories. Given the fact that diagonal models fail to outperform even with twice or three times as many mixture components, the results seem to suggest the pay-off is really from a better model for the residual errors rather than the trajectories themselves. This conjecture will be verified by further relaxing the assumptions made on the residual error in Sec. 3. One possible direction is to use mixture distributions to model the residual error.

Finally, the effectiveness of phone-pair segments and their consistently better performance than generalized triphone units are also interesting. In addition to phone-pairs being longer, there might be other explanations here. First, phone-pairs are designed to capture the transition trajectories across phone boundaries. Secondly, the overlapped evaluation framework illustrated in Figure 1 and 2 is able to minimize the impact of discontinuities for phone-pair SM's. A quick glance at the LVCSR outcomes does render the impression that recognition errors are correlated to triphone (monophone) back-off rates. We wish to explore context dependent phone-pair models in the future to further improve the system.

7. REFERENCES

[1] Ostendorf M., Digalakis V., and Kimball O., "From HMM's to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition", *IEEE Trans. on SAP, Sep. 1996.*

[2] Gish H. and Ng K. "A segmental Speech Model With Applications to Word Spotting", *ICASSP-1993.*

[3] Holmes W. and Russell M., "Probabilistic-Trajectory Segmental HMMs", *Computer Speech and Language, pp. 3-37, 1999.*

[4] Deng L., Aksmanovic M., Sun X., and Wu J., "Speech Recognition Using Hidden Markov Models with Polynomial Regression Functions as Nonstationary States", *IEEE Trans. on SAP, pp. 507-520, Oct. 1994.*

[5] Goldberger J., Burshtein D., and Franco H., "Segmental Modeling Using a Continuous Mixture of Nonparametric Models", *IEEE Trans on SAP, pp. 262-271, May 1999.*

[6] Siu M., Iyer R., Gish H., and Quillen C., "Parametric Trajectory Mixtures for LVCSR" *ICSLP-1998.*

[7] Iyer R., Gish H., Siu M., Zavaliagkos and Matsoukas S. "Hidden Markov Models for Trajectory Modeling", *ICSLP-1998.*

[8] Huang X., Acero A., Alleva F., Hwang M.Y., Jiang L. and Mahajan M. "Microsoft Windows Highly Intelligent Speech Recognizer: Whisper". *ICASSP-1995.*

[9] Wang K. and Goblirsch D., "Extracting Dynamic Features Using the Stochastic Matching Pursuit Algorithm for Speech Event Detection," *Proc. IEEE Workshop on ASRU, pp.132-139, 1997.*

[10] Gauvain J. and Lee C. "Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains" *IEEE Trans. on SAP, pp. 291-298, Apr. 1994.*