Segmental Modeling Using a Continuous Mixture of Non-parametric Models

Jacob Goldberger David Burshtein
Tel-Aviv University, Israel
jacob,burstyn@eng.tau.ac.il

Horacio Franco
SRI International, CA, USA
hef@speech.sri.com

10.21437/Eurospeech.1997-22

Abstract

The aim of the research described in this paper is to overcome the modeling limitation of conventional hidden Markov models. We present a segmental model that consists of two elements. The first is a nonparametric representation of both the mean and variance trajectories, which describes the local dynamics. The second element is some parameterized transformation (e.g., random shift) of the trajectory that is global to the segment and models long-term variations such as speaker identity.

Introduction

Speech sounds are produced by a time-varying dynamic system. Consequently, speech signals are highly correlated and nonstationary. In spite of this fact, in most implementations of hidden Markov models (HMMs) to speech recognition, the assumption that successive observations in a state are independent and identically distributed is inherent to the model. These limitations of the HMM are due to the fact that the HMM is a frame-based approach. An alternative approach is segmental modeling, where the basic modeling unit is not a frame but a phonetic unit. This family of models relaxes the assumptions of both stationarity and independence within a state, typical of standard HMMs. Deng *et al.* [1] used a regression polynomial function of time to model the trajectory of the mean in each state. A nonparametric description of the mean trajectory was suggested by Ghitza and Sondhi [4]. More recently, Kimball [6], suggested an approach that models each segment by a discrete mixture of nonparametric mean trajectories. In this paper we propose a new random segmental model. The main idea of random models is to consider the mean trajectory not as fixed parameters but as a random variable that is sampled on each state arrival. Russell and Holmes [5] proposed a random extension of the model suggested by Deng, by assuming a parametric segmental model with random coefficients. We suggest here a random nonparametric approach. Our proposed model is compared to alternative segment models by using a triphone recognition task. In addition, we present recognition results on a large vocabulary task.

Random Nonparametric Models

In this section we present a new segmental model which is composed of two elements. The first element is a nonparametric representation of the mean and variance trajectories, and the second is some parameterized transformation (e.g., random shift) of the trajectory that is global to the segment. The mean trajectory curve is represented using a nonparametric description. That is to say, instead of using a polynomial or some other parametric description, the curve is represented by specifying a list of sampled points along the curve. More precisely, we assume that each segment may be represented by a left to right HMM structure, such that each HMM state is represented by a single Gaussian HMM. The sequence of mean values of the HMM state sequence constitutes a template of the mean trajectory. Likewise, the sequence of variances of the HMM state sequence constitutes a template of the variance trajectory. Time warping of the template trajectory is made possible by controlling the state sequence of the HMM (e.g., companding may be realized by rapid transitions out of states). The second element of the model is a parameterized transformation of the trajectory, global to the entire segment. Let the state sequence of some given segment realization be denoted by $s = (s_1, s_2, \ldots, s_T)$, and let the corresponding observation vector sequence be denoted by $x = (x_1, x_2, \ldots, x_T)$. To simplify notation it will be assumed that all observations are scalars. This assumption is not necessary. We assume the following model:

$$x_t = T_a\left(\mu(s_t), \sigma(s_t), t\right)$$

where $\mu(s_t)$ and $\sigma(s_t)$ are the mean and variance associated with state s_t, and $T_a(\cdot)$ is some random transformation indexed by a. a is a random variable that is chosen once per segment realization. The transformation that we focus on in this paper, is a random displacement of the mean trajectory. In that

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

case $T_a(\mu,\sigma,t) = \mu + a + \epsilon_t(\sigma)$. Hence,

$$x_t = \mu(s_t) + a + \epsilon_t(\sigma(s_t)) \qquad (1)$$

Here, a is a zero mean, normal random variable, sampled once per segment, that represents the global displacement of the current segment realization. $\epsilon_t(\sigma)$ is a zero mean, Gaussian random variable.

$$a \sim N(0,\sigma_a^2) \quad , \quad \epsilon_t(\sigma) \sim N(0,\sigma^2)$$

The effect of the displacement variable may be interpreted as a continuous mixture of parallel curves that represent the mean trajectory along the segment. The distribution of a is the continuous segmental analog to the mixture coefficients in standard HMM. That is to say, in standard HMM a discrete mixture component is chosen once per frame, that is, it is a frame-based approach. In a random segmental model, however, a continuous mixture component is chosen once per segment realization.

Our model was motivated by extensive examination of segment data realizations. In Fig. 1, several realizations of the first cepstral coefficient in the triphone ih-s-ow are presented (The database used was the speaker-independent, large-vocabulary, Wall Street Journal (WSJ) corpus [2]). The data is presented after nonlinear time warping of the segment realizations, so as to achieve time alignment between the various realizations.

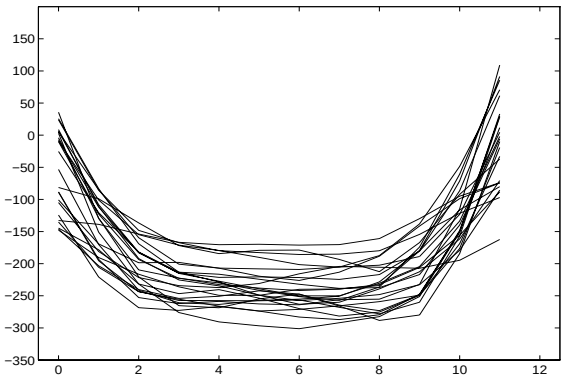

Figure 1: Data after nonlinear time warping

We now present recognition and training algorithms for the new proposed model. The input to the recognition algorithm is a segment realization. The output of the algorithm is the identity of the segment. Exact computation of the true probability $f(x)$ is not feasible. As an alternative, we propose using $\max_s f(x,s)$, which is an analog of the standard approximation of maximum likelihood (ML) word estimation by ML sequence estimation (Viterbi decoding). The following is an iterative algorithm to evaluate $\max_s f(x,s) = f(x,\hat{s})$ numerically:

1. Initialization: $\hat{a} = 0$.
2. Compute $\hat{s} = \arg\max_s f(x,s,\hat{a})$ by applying Viterbi segmentation on the data after displacement elimination (i.e., $x_1 - \hat{a}, x_2 - \hat{a}, \ldots, x_T - \hat{a}$).
3. Compute $\hat{a} = \arg\max_a f(x,\hat{s},a)$:

$$\hat{a} = \frac{\sum_t \frac{1}{\sigma^2(\hat{s}_t)}\left(x_t - \mu(\hat{s}_t)\right)}{\sum_t \frac{1}{\sigma^2(\hat{s}_t)} + \frac{1}{\sigma_a^2}} \qquad (2)$$

4. Repeat stages 2 and 3 until convergence.
5. Compute:

$$f(x,\hat{s}) = \left(\frac{1}{\sigma_a^2} + \sum_t \frac{1}{\sigma^2(\hat{s}_t)}\right)^{-\frac{1}{2}} \sqrt{2\pi} f(x,\hat{s},\hat{a}) \quad (3)$$

The proposed training algorithm is a combination of the algorithm above and the well-known Baum-Welch training procedure. Given a sequence of N segment data realizations $x_1, x_2, \ldots, x_N$, denote by a^i the segmental mixture coefficient of x_i. Training consists of the following iterative steps:

1. Initialization: $a_i = 0 \quad i = 1,2,\ldots,N$
2. Apply the Baum-Welch algorithm to $x_i - a_i$ to obtain a new set of segment template parameters (state means, variances, and transition probabilities). Then apply the Viterbi algorithm to compute state segmentation $\hat{s_i}$.
3. Apply equation (2) to obtain:
   $a_i = \arg\max_a f(x^i, \hat{s_i}, a)$.
4. Given $a_1, a_2, \ldots, a_N$, update the variance of the random displacement, a:

$$\sigma_a^2 = \frac{1}{N}\sum_{i=1}^{N} (a_i)^2$$

5. Repeat stages 2 through 4 until convergence.

A major decision that needs to be made concerns the number of states used in our model. On the one hand, trajectory descriptions with large numbers of states are more accurate. On the other hand, when a large number of states are used, the training algorithm needs to estimate a large number of parameters. Hence, in that case, it is essential to properly initialize the training algorithm. Otherwise, the algorithm does not produce meaningful results.

The following initialization algorithm is proposed.

1. Given the segment data realizations, $x_1, \ldots, x_N$, an initial segment template is determined. The length, M, of this template is set equal to the average length of the given segment realizations. Then each segment realization is linearly time warped to size M. Finally, the initial segment template is set to the mean of these linearly time-warped segment realizations.

2. A dynamic time warping (DTW) routine is used to time align each segment realization x^i against the template segment.

3. The time-aligned segment realizations are averaged together to obtain a new template.

4. Stages 2 and 3 are repeated as much as required. Typically, two iterations are sufficient.

5. Finally, M vectors of means and variances of the HMM states, constituting the initial template, are obtained by averaging the last version of time-aligned segment data realizations.

Note that the initialization routine does not employ random displacement modeling.

The DTW forces continuity constraints on the mean trajectory. Viterbi decoding does not incorporate such constraints, and thus does not produce reliable initialization.

The recognition and training algorithms described here are useful for re-scoring an N-best list. Note that because mean trajectory time warping is allowed, segmentation inaccuracies at the previous stage can be tolerated.

Experimental Results

We evaluated our model by using the ARPA large-vocabulary, speaker-independent, continuous-speech, Wall Street Journal (WSJ) corpus [2]. Experiments were conducted with DECIPHER, SRI's continuous speech recognition system [3].

Our model was implemented using the N-best re-scoring paradigm, by re-scoring the list of the N-best sentence hypotheses generated by DECIPHER. Context-dependent phonetic models were used. A segmental model was constructed for each triphone appearing in the training data set. The test set consisted of 200 sentences. Table 1 compares acoustic performance. Table 2 address the issue of adding the segmental model as another knowledge source to a complete recognition system.

model	word error
HMM acoustics	22.1
segmental acoustics	21.4

Table 1: Word error rate results without language model.

Tables 1 and 2 show that the new model is comparable to a state-of-the-art HMM system, with sophisticated tying of parameters. To probe the new model further and to compare it to alternative models, we carried out several triphone recognition experiments. Context-dependent phonetic units were chosen because, in that case, there are fewer discrepancies between utterances. Hence, in practice, this is usually the case of interest.

model	word error
HMM acoustics + linguistics	8.1
HMM acoustic + linguistics + segmental acoustics	7.8

Table 2: Word error rate results with language model.

Table 3 presents recognition results for some frequently occurring triphone contexts. The first data row indicates the number of triphone occurrences for each context. Half of the occurrences were used to train each model. The other half were used to test the models. There were six triphones in the first context (s[k]ih, s[l]ih, s[m]ih, s[p]ih, s[t]ih and s[w]ih), five triphones in the second context (n[ay]t, n[eh]t, n[ey]t, n[ih]t and n[ow]t), five triphones in the third context (aa[k]t, aa[n]t, aa[p]t, aa[r]t and aa[s]t), ten triphones in the fourth context (ih[b]eh, ih[d]eh, ih[f]eh, ih[jh]eh, ih[l]eh, ih[m]eh, ih[p]eh, ih[r]eh, ih[s]eh and ih[v]eh), and seven triphones in the fifth context (g[aa]t, g[ae]t, g[ah]t, g[ax]t, g[eh]t, g[ey]t and g[ih]t).

The following models were examined:

1. An HMM having a mixture of Gaussians. Such a model with s states and m mixtures is denoted by HMM(s,m).

2. A segmental polynomial model [1] with deterministic coefficients. Such a model with s states and a polynomial of order K describing the mean trajectory of each state is denoted by POLY(s,K).

3. A segmental random polynomial model [7] with multinormal coefficients. Such a model with s states and a polynomial of order K describing the mean trajectory of each state is denoted by POLYRND(s,K).

4. The new proposed model with random displacement modeling. Such a model with s states is denoted by NPRMDISP(s).

5. The new proposed model without random displacement modeling, that is, a standard nonparametric model. Such a model with s states is denoted by NPRM(s).

As can be seen, in four out of the five contexts presented, global random displacement, nonparametric modeling (NPRMDISP) is preferable to standard nonparametric segmental modeling (NPRM). The new model also compares favorably with the other models that were examined.

	s[·]ih	n[·]t	aa[·]t	ih[·]eh	g[·]t
#	1088	740	2263	1619	662
HMM(3,3)	90.7	85.2	96.6	89.3	64.1
POLY(3,2)	89.0	82.7	95.9	87.5	66.8
POLYRND(3,1)	89.6	79.2	96.3	87.4	64.4
NPRM(9)	90.7	78.7	94.5	89.9	58.7
NPRMDISP(9)	91.6	85.4	96.5	87.9	67.1

Table 3: Triphone recognition rate results

The experiments summarized in Table 3 were repeated for many other frequently occurring triphone contexts. For most triphone contexts that were examined, random displacement modeling improved the standard nonparametric model. Nevertheless, in many other cases, random displacement modeling decreased the recognition rate. Hence, for some of the triphones, a standard nonparametric model (i.e., a degenerated displacement model that employs fixed zero displacement) is expected to be preferable. On the other hand, we observed that a random displacement model always assigns higher likelihood values to previously unseen data, and hence has an improved prediction capability. Therefore, the maximum likelihood criterion cannot be used to decide when (i.e., for which triphones) the random displacement model should be avoided.

Conclusions

We presented a new model that is a continuous mixture of segment trajectories. This model is composed of two elements. The first element is a nonparametric representation of the mean and variance trajectories, and the second is some parameterized transformation of the trajectory that is global to the entire segment. This transformation adapts the general model to a specific segment realization, and may, for example, account for different speech styles. We then focused on a particular transformation that applies a random displacement to the mean trajectory. The model was compared to alternative segment models on a triphone recognition task. The model improves segment modeling in the sense that it improves the prediction of previously unseen data. Our triphone recognition experiments show model efficacy for most of the contexts examined when compared with a standard nonparametric model without global displacement modeling.

The results presented using this model suggest several topics for future study. First, other global trajectory transformations need to be examined. Second, we have seen that a global, random displacement transformation always improves the ability of a nonparametric model to predict previously unseen data. However, the new model was not always superior in the triphone recognition experiments. The maximum likelihood criterion cannot be used to decide for which triphones random displacement modeling should degenerate to a fixed zero displacement. Other criteria need to be investigated to successfully implement a combined model, for which some of the triphones employ such degenerated transformation.

Acknowledgement

We gratefully acknowledge partial support for this work from ARPA through Office of Naval Research Contract N00014-94-C-0181.

References

[1] L. Deng, M. Aksmanovic, D. Sun and J. Wu, "Speech recognition using hidden Markov models with polynomial regression functions as non stationary states", *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 507-520, 1994.

[2] G. Doddington, "CSR Corpus Development", *Proc. ARPA Workshop on Spoken Language Technology*, Feb. 1992.

[3] V. V. Digalakis, P. Monaco and H. Murveit, "Genones: generalized mixture tying in continuous hidden Markov model-based speech recognizers", *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 281-289, 1996.

[4] O. Ghitza and M. Sondhi, "Hidden Markov models with templates as non-stationary states: an application to speech recognition", *Computer Speech and Language*, vol. 7, pp. 101-119, 1993.

[5] W. Holmes and M. Russell, "Experimental evaluation of segmental HMMs", *Proc. Int. Conf. Acoust., Speech, Signal Processing*, pp. 536-539, 1995.

[6] O. Kimball, "Segment modeling alternatives for continuous speech recognition", Ph.D thesis, Boston University, 1994.

[7] M. Russell, "A segmental HMM for speech pattern modeling", *Proc. Int. Conf. Acoust., Speech, Signal Processing*, pp. 499-502, 1993.