Segmental Modeling Using a Continuous Mixture of Non-Parametric Models

Jacob Goldberger David Burshtein Horacio Franco

Tel-Aviv University, Israel SRI International, CA, USA

jacob,burstyn@eng.tau.ac.il hef@speech.sri.com

EDICS Categories: SA 1.6, SA 1.6.3.

Abstract — A major limitation of hidden Markov model (HMM)-based automatic speech recognition is the inherent assumption that successive observations within a state are independent and identically distributed (IID). The IID assumption is reasonable for some of the states (e.g., a state that corresponds to a steady state vowel). However, most states clearly violate this assumption (e.g. states corresponding to vowel-consonant transition, diphthongs, etc.) and are in fact characterized by a highly correlated and non-stationary speech signal. In recent years, alternative models have been proposed, that attempt to describe the dynamics of the signal within a phonetic unit. The new approach is generally known by the name segmental modeling, since the speech signal is modeled on a segment level base and not on a frame base (such as HMM). We propose a family of new segmental models that are composed of two elements. The first element is a non-parametric 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 entire segment. The new model is in fact a continuous mixture of segment trajectories. We present recognition results on a large vocabulary task, and compare the model to alternative segment models on a triphone recognition task.

1 Introduction

Hidden Markov model (HMM)-based automatic speech recognition (ASR) has attracted a great deal of interest in the last two decades. The performance of speech recognizers that employ hidden Markov modeling has been shown to be superior compared to that of alternative recognition methods in a variety of real life applications, and in particular for speaker independent large vocabulary tasks.

The standard left to right HMM provides a technique for modeling the acoustic feature vector sequence, that represents some speech utterance, by a piecewise stationary process. The model assumes the existence of states, such that the observations are locally independent and identically distributed (IID) within each state. Several HMM variants were suggested. The simplest variant employs a discrete output probability distribution function (PDF) to describe the acoustic feature vector at each HMM state. A refinement of the above is obtained by replacing the discrete PDF with a continuous PDF, which is usually a mixture of Gaussians with a diagonal covariance matrix. All these variants share a common assumption that the probability of an acoustic vector in a particular state does not depend on the other vectors in that state. This simplifying assumption assures computationally efficient algorithms for system training (forward-backward) and recognition (Viterbi). It should be noted though, that these efficient algorithms can be extended to the case where the independence assumption between vectors in a state is replaced by a Markovian one [26].

The IID assumption is reasonable for some of the HMM states (e.g., states that corresponds to a steady state vowel in a user dependent system). However, most states clearly violate this assumption (e.g., states corresponding to vowel-consonant transition, diphthongs, etc.) and are in fact characterized by a highly correlated and non-stationary speech signal. The consequence is a reduced accuracy of speech acoustic modeling which is translated into reduced recognition rate. In recent years, alternative models, that attempt to overcome these difficulties were proposed [20].

These methods are usually known by the name segmental models.

In this paper we propose a family of new segmental models that are composed of two elements. The first element is a non-parametric 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 entire segment. The new model is in fact a continuous mixture of segment trajectories. The proposed model is compared to alternative segment models on equal terms, by using a triphone recognition task. In addition, we present recognition results on a large vocabulary task.

The organization of this paper is as follows. In the next section we indicate difficulties and limitations of the standard HMM. In section 3 we review some segmental HMMs that were previously suggested in the literature. Section 4 introduces the new segmental model. Section 5 presents experiments that were conducted with the new model. Section 6 concludes the paper.

2 Limitations of the standard HMM approach

The traditional formulation of a single Gaussian HMM (i.e., with mixture size equal to one) has been based on a piecewise constant fitting of the acoustic feature vector data sequence [21], [22]. This model assumes that the observation vectors within a state, $x_t, \quad t = 1, \ldots, T$ are generated according to

$$x_t = \mu + \epsilon_t$$

where μ is a state dependent parameter that represent the mean vector and ϵ_t is an additive, zero mean, white noise vector (i.e., its covariance matrix is diagonal) with state dependent variances. In Fig. 1 we present the sixth cepstral coefficient of the acoustic vector sequence of the word 'seven'. The smooth curve is obtained by empirical averaging of utterances of the word 'seven' in the TIDIGITS database [17]. This database consists of high quality connected digit string utterances by 225 adult female and male speakers. Before averaging, the utterances were time

aligned by using a non linear dynamic time warping transformation. In Fig. 1 we also present a five states HMM piecewise constant approximation to that curve. This approximation consists of the means of the most likely state sequence corresponding to the given curve.

The most common method of encoding the dependency between consecutive frames is to extend the feature vector to include the first and sometimes the second derivative of the static features. In Fig. 2 we present the same data that was presented in Fig. 1, for the first derivative of the sixth cepstral coefficient.

In a standard state-stationary HMM, the trajectories of the feature vectors are approximated by a piecewise constant function. Each region of constant value corresponds to an HMM state. As can be seen from Fig. 1 and Fig. 2, a piecewise constant function is usually a poor approximation for the mean trajectory. Another problem of this model arises from the fact that we attempt to model simultaneously a static feature vector (e.g. cepstrum) and its time derivative. To observe the disadvantage associated with this model, consider Fig. 3, that presents a feature vector that consists of two components. The first (static) component is a sine wave; the second is the time derivative of the first, hence a $\pi/2$ phase shifted sine wave. Fig. 3 also presents the optimal partitioning into states of these components, so as to achieve best piecewise constant approximation, and the resulting approximation. As can be seen, the state partitioning that is required for optimally approximating the mean trajectory of the first component is different from the one required for approximating the mean trajectory of the second. However, since the same state partitioning should be used for both components, this results in approximation of reduced quality.

To gain further insight, consider a speaker independent, mixture of Gaussians, HMM system. The improved performance of this system, compared to a single Gaussian HMM is usually attributed to the fact that mixtures help to improve the modeling of the true state distribution which is clearly non-Gaussian, and in this way improve the modeling of the variation between different speech styles.

We now suggest an alternative explanation to this phenomenon. Essentially, we assert that mixtures help to describe the non-stationary behavior of the feature vectors within a state. Assume that the trajectory of the mean in a state is changing from a_0 at the beginning of the state to a_1 at the end. By setting the mean values of three mixture components to a_0, $(a_0 + a_1)/2$ and a_1 we may obtain improved modeling of the mean trajectory by assigning these mixture values to the beginning, middle and final periods of duration at that state.

In Fig. 4 we present the sixth cepstral coefficient of the acoustic feature vector, and its HMM approximation (the mean values of the mixture terms at that state), for a word-based HMM with five states and three mixtures (the word shown is 'seven'). In Fig. 5 we present the same data for the time derivative of the sixth cepstral coefficient. These figures demonstrate how the model employs mixtures to track feature trajectories. As an example, consider the second state in Fig. 4. The three mixture components yield better approximation to the feature trajectory, compared to a single Gaussian HMM, since different mixture components may be used when approximating different portions of the trajectory.

To further assess the validity of the proposed explanation, we recorded the statistics of the transition between mixtures, within some given state, by using the sequence of most likely mixture at each frame. For that purpose, a word based HMM recognizer was used on the speaker independent TIDIGITS database. Each digit was modeled by a five state HMM, with three mixtures per state. The transition matrix of the mixtures at the second state of the word 'seven' is shown in Table 1. As can be seen, the IID within a state assumption is not valid in practice. In fact, there is a clear trend to choose the mixtures in a fixed order (the 3'rd mixture first, then the 2'nd mixture, and finally the 1'st mixture). Similar data is presented in Table 2, except that a sequence of consecutive self transitions is now counted only once. Tables 1 and 2 validate the explanation that we set above.

3 Review of Existing Segmental HMMs

The local IID within a state assumption of standard HMMs implies that we model the data on a frame-base level and ignore the continuous dynamics of the signal within a state. An alternative approach is segmental modeling, where the basic modeling unit is not a frame but a phonetic unit. This family of models relax both the stationarity and the independence within a state assumptions of standard HMMs. In this section we review major variants of segmental models. A more detailed survey of segmental models can be found in [20].

Deng *et al.* [1] used a regression polynomial function of time to model the trajectory of the mean in each state. A similar model was suggested by Gish and Ng [9] for a keywords spotting task. In that model, the observation vectors within a state are generated according to

$$x_t = \sum_{k=0}^{K} \mu_k t^k + \epsilon_t \tag{1}$$

such that t is set to zero at the beginning of the state and then incremented with each new incoming frame. μ_k are state dependent vector parameters, and ϵ_t is a zero mean Gaussian, with a state dependent diagonal covariance matrix. The case K=0 corresponds to standard HMM. This model assumes that the frames within a state are independently (although not identically) distributed.

Russell and Holmes [23], [12], [13], [14], and Gales and Young [6] [7] extended the model suggested by Deng, by assuming a parametric segmental model with random coefficients, that are sampled once per segment realization. Therefore, the mean trajectory is a stochastic process instead of a fixed parameter. More precisely, this model is defined by Eq. (1) and by the PDFs of $\mu_k \quad k = 0, 1, \ldots, K$ and ϵ_t. The resulting probabilistic model may be viewed as a double stage sampling process, where in the first stage we sample the model coefficients, μ_k, using their PDFs. In the second stage we create the observations x_t, by sampling along the parametric curve that was determined in the first stage. This sampling is carried out with the PDF of ϵ_t. Diagonal

covariance Gaussian PDFs are typically attributed to μ_k and ϵ_t. In addition, ϵ_t is assumed to have zero mean. The model parameters can be normalized according to the segment length in order to achieve better performance and to simplify the parameter estimation [10].

Kenny *et al.* [15] have used a state conditioned linear prediction coefficients (LPC) model to remove correlation between successive observation vectors, i.e. the observation vectors within a state are generated according to

$$x_t = \mu + \sum_{k=1}^{K} \beta_k x_{t-k} + \epsilon_t \tag{2}$$

where β_k are diagonal matrices, so that a LPC model applies to each component of the vector x_k. A disadvantage of the model is that it assumes stationarity within a state. The two approaches of [1] and [15] were unified and generalized in [2]. Digalakis [4] proposed a dynamical system model which generalizes the Gauss-Markov model (2) to a Kalman filter framework, by assuming noisy observations. The special case where the hidden Gauss-Markov process is assumed to be constant was named 'target state model'. The target state model is similar to the model proposed by Russell [23]. Therefore, the dynamical system model can also be considered a generalization of the hidden constant Gaussian mean (target state) model.

Several authors have proposed non parametric segment models. A major advantage of non parametric models is that they are not sensitive to the shape of the feature trajectory that needs to be approximated. Consequently, they are also not sensitive to the segment partitioning problem that was explained in section 2 and demonstrated in Fig. 3 for a horizontal line parametric approximation. On the other hand, non parametric models might require more data to train the model on, since they are less constrained than parametric models. The first non-parametric approach to a non-stationary state HMM was the stochastic segment model (SSM) suggested by Ostendorf and Roukos [18] in 1989. The SSM assigns a Gaussian distribution to the entire segment which is resampled to a fixed length. A non parametric approach to a non-stationary state HMM with an additional step of time warping was suggested by Ghitza and Sondhi [8].

In [8], the trajectory of the mean in a given state is set equal to that state realization in the training set whose DTW distance [24] from all other sequences in the ensemble is minimal. More recently, Kimball *et al.* [16], [20] suggested a non-parametric approach that models each segment by a discrete mixture of non-parametric mean trajectories.

Direct implementation of segmental models is typically computationally demanding. This is due to the fact that the exact beginning and ending points of the segment must be given in order to compute an acoustic score. The N-best paradigm [25] offers a solution to this problem, by using the following two stage recognition procedure. At the first stage, a standard HMM recognition system is used to produce a list of size N of best hypothesized candidate strings, with the associated acoustic segmentation of each hypothesis. At the second stage, a more informative segmental acoustic model is used to re-score these candidates. Essentially, the N-best paradigm takes advantage of the computational efficiency of standard HMM recognition.

4 Continuous Mixture of Non-Parametric Segmental Models

In this section we present a new segmental model which is composed of two elements. The first element is a non-parametric 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 entire segment. The mean trajectory curve is represented using a non-parametric description. That is to say, instead of using a polynomial or some other parametric description, as in [1], [2], [15], [4], [23], [12], [13], [14] and [7], the curve is represented by specifying a list of sampled points along the curve. In that sense the proposed model is similar to the SSM suggested in [18]. 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., contracting may be realized by rapid transitions out of states). The second element of the model is a parameterized transformation of the trajectory, that is 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 $x = (x_1, x_2, \ldots, x_T)$. The dimension of x_i is L. We assume the following model for the observation vector, x_t, at time t:

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

where $\mu(s_t) = (\mu_1(s_t), \ldots, \mu_L(s_t))$ and $\sigma(s_t) = (\sigma_1(s_t), \ldots, \sigma_L(s_t))$ are the mean and variance vectors at 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 case $T_a(\mu, \sigma, t) = \mu + a + \epsilon_t(\sigma)$. Hence,

$$x_t = \mu(s_t) + a + \epsilon_t(\sigma(s_t)) \tag{3}$$

Here, a is a zero mean, multi-normal random variable, sampled once per segment, that represents the global displacement of the current segment realization:

$$a \sim N(0, \mathrm{diag}(\alpha_1^2, \ldots, \alpha_L^2))$$

where $\mathrm{diag}(\alpha_1^2, \ldots, \alpha_L^2)$ denotes a diagonal matrix with diagonal elements $\alpha_1^2, \ldots, \alpha_L^2$. $\epsilon_t(\sigma)$ is a zero mean, diagonal covariance Gaussian:

$$\epsilon_t(\sigma) \sim N(0, \mathrm{diag}(\sigma_1^2, \ldots, \sigma_L^2))$$

The effect of the displacement variable may be interpreted as a continuous mixture of parallel curves that represents 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, i.e., it is a frame based approach,

while in a random segmental model, a continuous mixture component is chosen once per segment realization.

The proposed model (3) is similar to the segmental models suggested in [23] and [7]. In these references, however, the approach is parametric, while our approach is non-parametric and allows time warping of the mean trajectory. The flexibility that is gained by allowing time warping can significantly improve the modeling capability of the template trajectory. In particular, the model tolerates inaccuracies in the hypothesized phone boundaries. Non-parametric modeling of the mean trajectory (without a global displacement element) were already proposed in [8] and [11]. However, our estimation procedure seems to be more robust, since it utilizes averaging of the various realizations. Digalakis and Ostendorf [4], [19] also proposed a non-parametric segment trajectory model. A major difference between our proposed approach and these references is that we use dynamic time warping instead of the linear time warping used in [4] and [19]. The origin of non-unique mean trajectory modeling can be attributed to the traditional mixture HMM. Kimball and Ostendorf [16], [20] described a segmental model that consists of discrete mixtures of trajectories.

Our model was motivated by extensive examination of segment data realizations. In Fig. 6, 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 [3]). Fig. 7 presents the same data after non linear time warping of the segment realizations, so as to achieve time alignment between the various realizations. Fig. 8 also presents time aligned segment realizations, but with an additional stage of displacement removal (the estimated displacement, a of each segment is obtained using an algorithm that is presented below). It is clearly seen, that the variance of the trajectories in Fig. 8 is smaller than the corresponding variance in Fig. 7. Hence, in this case, the incorporation of the global displacement term yields improved modeling. This observation is the motivation to the model (3).

These figures reflect the fact that the speech signal is produced by a continuously varying

physical system (the vocal track). There is a smooth local dynamic along the sojourn in the phonetic unit. We can also observe from these figures that when we fix a context there is no significant variability among different realizations of a phonetic unit. In other words, the phone does not exhibit several distinct trajectories. The observed difference can be described as a shift from the mean trajectory that is global to the entire segment. This phenomenon can be explained on two levels. First, the shift reflects the speaker personal style which is dependent upon the shape of vocal organs, gender, age, dialect etc. From a more local point of view the size of the shift depends on the ending point of the previous phoneme and the starting point of the next. This is one way of describing the co-articulation effect. The vocal articulators move from the position necessary for articulation of the previous phone towards the position required for the next phone, via the position needed for the current phone. During fluent speech, phones are sometimes not fully articulated. Consequently, due to the continuous nature of the speech signal, the mean trajectory is shifted. In the next section we shall formulate these intuitive observations into a probabilistic model.

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. The optimal maximum likelihood (ML) solution to this problem is to make a decision on the segment identity $\hat{p}$, based on the likelihood of the segment data x, i.e.

$$\hat{p} = \arg\max_{p \in \mathcal{P}} f(x) \tag{4}$$

where $\mathcal{P}$ is the set of candidate segments, and $f(x)$ is the density of x under the assumption that the segment identity is p (For simplicity we do not indicate the dependence of $f(x)$ on p. A more precise notation for the density of X at x is $f_X(x)$. However, we are using the standard shortened notation, $f(x)$. This convention is also used elsewhere, e.g. the joint density of X, S and A at x, s, a is denoted by $f(x, s, a)$). $f(x)$ is given by:

$$f(x) = \int_a \sum_s f(x, s, a) da$$

A disadvantage of the ML approach is that the computation of $f(x)$ is very complicated. As an alternative we propose the following segment recognition criterion:

$$\hat{p} = \arg \max_{p \in \mathcal{P}} \left\{ \max_s f(x, s) \right\} \tag{5}$$

The approximation of (4) by (5) is similar to the standard approximation of ML word estimation by ML sequence estimation (Viterbi decoding). We now present 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 standard 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 $\hat{a} = \arg \max_a f(x, \hat{s}, a)$. In the Appendix we obtain the following expression for $\hat{a}_j$, the j-th component of $\hat{a}$.

$$\hat{a}_j = \frac{\sum_{t=1}^{T} \frac{1}{\sigma_j^2(\hat{s}_t)} \left(x_{t,j} - \mu_j(\hat{s}_t) \right)}{\sum_{j=1}^{T} \frac{1}{\sigma_j^2(\hat{s}_t)} + \frac{1}{\alpha_j^2}} \qquad j = 1, 2, \ldots, L \tag{6}$$

   where $x_{t,j}$, $\mu_j(\hat{s}_t)$ and $\sigma_j(\hat{s}_t)$ are the j-th components of the vectors x_t, $\mu(\hat{s}_t)$ and $\sigma(\hat{s}_t)$, respectively.

4. Repeat 2 and 3 until convergence.

5.

$$f(x, \hat{s}) = \left(\frac{1}{\sqrt{2\pi}} \right)^{L \cdot T} f(\hat{s}) \prod_{j=1}^{L} \left\{ \frac{1}{\alpha_j \prod_{t=1}^{T} \sigma_j(\hat{s}_t)} \frac{1}{\sqrt{\frac{1}{\alpha_j^2} + \sum_{t=1}^{T} \frac{1}{\sigma_j^2(\hat{s}_t)}}} \right\} \exp \left\{ -\frac{1}{2} \sum_{j=1}^{L} k_j(x, \hat{s}) \right\} \tag{7}$$

where

$$k_j(x,s) = \sum_{t=1}^{T} \frac{1}{\sigma_j^2(s_t)} \left(x_{t,j} - \mu_j(s_t)\right)^2 - \frac{\left(\sum_{t=1}^{T} \frac{1}{\sigma_j^2(s_t)} \left(x_{t,j} - \mu_j(s_t)\right)\right)^2}{\frac{1}{\alpha_j^2} + \sum_{t=1}^{T} \frac{1}{\sigma_j^2(s_t)}}$$

Eq. (7) is derived in the Appendix.

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 $x^1, x^2, \ldots, x^N$, where $x^i = (x_1^i, x_2^i, \ldots, x_{T_i}^i)$, denote by $a^i = (a_1^i, \ldots, a_L^i)$, the segmental mixture coefficient vector of x^i. Training consists of the following iterative steps:

1. Initialization: $a^i = 0 \quad i = 1, 2, \ldots, N \quad , \quad \alpha_j^i = \infty \quad i = 1, \ldots, N, \quad j = 1, \ldots, L.$

2. Apply the Baum-Welch algorithm to $x_1^i - a^i, x_2^i - a^i, \ldots, x_{T_i}^i - a^i$, in order to obtain a new set of segment template parameters (state means, variances, and transition probabilities). We then apply the Viterbi algorithm in order to compute state segmentation $\hat{s}$.

3. Apply the previous iterative algorithm to obtain $a^i = \arg\max_a f(x^i, \hat{s}, a)$.

4. Given $a^1, a^2, \ldots, a^N$, update the variance of the random displacement, a:

$$\alpha_j^2 = \frac{1}{N} \sum_{i=1}^{N} \left(a_j^i\right)^2 \quad j = 1, \ldots, L$$

5. Repeat 2-4 until convergence.

A major decision that needs to be made concerns the number of states that are used in our model. On the one hand, trajectory descriptions with large number 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 that was described above. Otherwise, the algorithm does not produce meaningful results. This problem is avoided in Kimball *et al.* [16], [20], since a relatively small number of states (typically five states per segment) is used.

The following initialization algorithm is proposed.

1. Given the segment data realizations, $x^1, x^2, \ldots, x^N$, 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 in order 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, that constitute 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 routine that we used specifies the local constraint that no more than two adjacent template frames can be mapped to the same segment realization frame, and vice versa (no more than two adjacent segment realization frames can be mapped to the same template frame). In addition, the DTW routine specifies the standard global constraint that the grid region of matching frames is limited to a band diagonal region. These constraints limit the amount of permitted time companding and expanding. Standard Viterbi decoding does not incorporate such constraints, and thus does not produce reliable initialization.

The recognition and training algorithms that were described above are useful for re-scoring an N-best list. Note that due to the fact that mean trajectory time warping is allowed, segmentation

inaccuracies at the previous stage can be tolerated.

5 Experimental Results

We evaluated the model presented in the previous section using the ARPA, large vocabulary, speaker independent, continuous speech, Wall Street Journal (WSJ) corpus [3]. We used the 206 male sentences of the development set for the 1994 ARPA sponsored evaluations using a 5K close vocabulary language model.

Experiments were conducted with DECIPHER, SRI's continuous speech recognition system [5]. The recognizer was configured with a front end that outputs a 39-dimensional vector. The first components of the vector consists of 12 cepstral coefficients and an energy term. The other components of the feature vector are the first and second time derivatives of the first 13 components.

Our model was implemented using the N-best re-scoring paradigm, by re-scoring the list of the N-best sentence hypotheses generated by the DECIPHER. N-best lists were generated using a 5K close vocabulary bigram language model and HMM genone [5] gender-dependent models trained on the male set of the WSJ corpus. When rescoring the N-best lists we used a 5K close vocabulary trigram language model. A segmental model was constructed for each triphone that appears in the training data set. In Table 3 we present the word error rate of standard HMM, as implemented in the DECIPHER, and the word error rate after re-scoring the N-best list using the segmental model. In that case, language modeling was not incorporated. In Table 4 we compare the performance of HMM acoustics to segmental acoustics when linguistic information is incorporated. Although in Table 3 segmental acoustics is superior to HMM acoustics, in Table 4 HMM acoustics is seen to be superior. This can be attributed to the much higher complexity of the HMM (32 continuous mixtures). In addition to that, the language model was tuned to HMM acoustic scoring. In Table 4 we also show the decrease in the word error rate when we add the

segmental model to the HMM as another knowledge source, and linearly combine the two scores. As can be seen, segmental acoustics adds some extra information.

Tables 3 and 4 show that the new model is comparable to 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 since in that case there is less variability between utterances. Hence, in practice, this is usually the case of interest.

In Table 5 we present recognition results for some frequently occurring triphone contexts. This is not a conventional recognition experiment, but it provides a way to assess the relative discriminative power of different acoustic models. 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 was 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 forth 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 models examined were:

1. Mixture of Gaussians HMM. Such model, with s states and m mixtures is denoted by HMM(s,m).

2. A segmental polynomial model (1) with deterministic μ_k parameters. Such 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 polynomial model (1) with multi-normal μ_k parameters. Such 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 model with s states is denoted by NPRMDISP(s).

5. The new proposed model without random displacement modeling, i.e. a standard non-parametric model. Such model with s states is denoted by NPRM(s).

To implement model (1) (both for the case where μ_k are deterministic parameters, and for the case where they are random variables), all possible state partitions were considered for each utterance that needs to be recognized.

The values of s, m and K in the various models were set such that they all have (approximately) the same number of free parameters.

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

The experiments that are summarized in Table 5 were repeated for many other frequently occurring triphone contexts. For most triphone contexts that were examined, random displacement modeling improved the standard non-parametric model. Nevertheless, in many other cases, random displacement modeling decreased the recognition rate. Hence, for some of the triphones, a standard non-parametric 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 in order to decide when the random displacement model should be degenerated.

6 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 non-parametric 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 benefit to the new model for most of the contexts examined, compared to a standard non-parametric model without global displacement modeling.

Several avenues of future research suggest themselves. First, other global trajectory transformations need to be examined. One possibility is to consider transformations that control the sharpness of picks and valleys of the mean trajectory. In that case, the model accounts for segment utterances with varying degrees of smoothness.

Second, we have seen that a global, random displacement transformation always improves the ability of a non-parametric 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 in order to decide for which triphones random displacement modeling should degenerate to a fixed zero displacement. Other criteria need to be investigated in order to successfully implement a combined model, for which some of the triphones employ such degenerated transformation.

Appendix

We derive Eqs. (6) and (7). Denote by $f(x,s,a)$ the joint density function of the data x, the states sequence s, and the shift a. Similarly $f(x,s)$ is the joint density of x and s, and $f(s)$ is the marginal density of s. We first derive the optimal shift, denoted by $\hat{a}$, given x and s.

$$f(x,s,a) = f(s)\prod_{j=1}^{L}\left\{\frac{1}{\sqrt{2\pi}\alpha_j}\exp\left\{-\frac{a_j^2}{2\alpha_j^2}\right\}\prod_{t=1}^{T}\frac{1}{\sqrt{2\pi}\sigma_j(s_t)}\exp\left\{-\frac{(x_{t,j}-a_j-\mu_j(s_t))^2}{2\sigma_j^2(s_t)}\right\}\right\}$$

$$\frac{\partial \log f(x,s,a)}{\partial a_j} = -\frac{a_j}{\alpha_j^2} + \sum_{t=1}^{T}\frac{1}{\sigma_j^2(s_t)}\left(x_{t,j}-a_j-\mu_j(s_t)\right) = 0$$

Therefore the maximum likelihood shift is :

$$\hat{a}_j = \arg\max_{a_j} f(x,s,a) = \frac{\sum_{t=1}^{T}\frac{1}{\sigma_j^2(s_t)}\left(x_{t,j}-\mu_j(s_t)\right)}{\sum_{t=1}^{T}\frac{1}{\sigma_j^2(s_t)}+\frac{1}{\alpha_j^2}} \quad j=1,2,\ldots,L$$

Next we derive a closed form expression to $f(x,s)$.

$$\begin{aligned} f(x,s) &= \int_a f(x,s,a)da \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^{L(T+1)} f(s)\prod_{j=1}^{L}\left\{\frac{1}{\alpha_j\prod_{t=1}^{T}\sigma_j(s_t)}\int_{a_j}\exp\left\{-\frac{1}{2}g_j(x,s,a_j)\right\}da_j\right\} \end{aligned}$$

where

$$g_j(x,s,a_j) = \frac{a_j^2}{\alpha_j^2} + \sum_{t=1}^{T}\frac{1}{\sigma_j^2(s_t)}\left(x_{t,j}-a_j-\mu_j(s_t)\right)^2$$

Now,

$$\begin{aligned} g_j(x,s,a_j) &= a_j^2\left(\frac{1}{\alpha_j^2}+\sum_{t=1}^{T}\frac{1}{\sigma_j^2(s_t)}\right) - 2a_j\sum_{t=1}^{T}\frac{1}{\sigma_j^2(s_t)}\left(x_{t,j}-\mu_j(s_t)\right) + \sum_{t=1}^{T}\frac{1}{\sigma_j^2(s_t)}\left(x_{t,j}-\mu_j(s_t)\right)^2 \\ &= \left(\frac{1}{\alpha_j^2}+\sum_{t=1}^{T}\frac{1}{\sigma_j^2(s_t)}\right)\left(a_j - \frac{\sum_{t=1}^{T}\frac{1}{\sigma_j^2(s_t)}\left(x_{t,j}-\mu_j(s_t)\right)}{\frac{1}{\alpha_j^2}+\sum_{j=1}^{T}\frac{1}{\sigma_j^2(s_t)}}\right)^2 + \\ &\quad \sum_{t=1}^{T}\frac{1}{\sigma_j^2(s_t)}\left(x_{t,j}-\mu_j(s_t)\right)^2 - \frac{\left(\sum_{t=1}^{T}\frac{1}{\sigma_j^2(s_t)}\left(x_{t,j}-\mu_j(s_t)\right)\right)^2}{\frac{1}{\alpha_j^2}+\sum_{t=1}^{T}\frac{1}{\sigma_j^2(s_t)}} \\ &= \left(\frac{1}{\alpha_j^2}+\sum_{t=1}^{T}\frac{1}{\sigma_j^2(s_t)}\right)\left(a_j-\hat{a}_j\right)^2 + k_j(x,s) \end{aligned}$$

where

$$k_j(x,s) = \sum_{t=1}^{T}\frac{1}{\sigma_j^2(s_t)}\left(x_{t,j}-\mu_j(s_t)\right)^2 - \frac{\left(\sum_{t=1}^{T}\frac{1}{\sigma_j^2(s_t)}\left(x_{t,j}-\mu_j(s_t)\right)\right)^2}{\frac{1}{\alpha_j^2}+\sum_{t=1}^{T}\frac{1}{\sigma_j^2(s_t)}}$$

(Note: $k_j(x,s) = g_j(x,s,\hat{a}_j)$). Hence,

$$f(x,s) = \left(\frac{1}{\sqrt{2\pi}}\right)^{L\cdot T} f(s) \prod_{j=1}^{L} \left\{ \frac{1}{\alpha_j \prod_{t=1}^{T} \sigma_j(s_t)} \frac{1}{\sqrt{\frac{1}{\alpha_j^2} + \sum_{t=1}^{T} \frac{1}{\sigma_j^2(s_t)}}} \right\} \exp\left\{ -\frac{1}{2} \sum_{j=1}^{L} k_j(x,s) \right\}$$

Acknowledgment

We gratefully acknowledge partial support for this work from DARPA 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] L. Deng and C. Rathinavelu, "A Markov model containing state-conditioned second order non-stationary: application to speech recognition", *Computer Speech and Language*, vol. 9, pp. 63-86, 1995.

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

[4] V. V. Digalakis, "Segment-based stochastic models of spectral dynamics for continuous speech recognition", E.C.S Department, Ph.D Thesis, Boston University, 1992.

[5] 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.

[6] M. Gales and S. J. Young, "The theory of segmental hidden Markov models", Technical Report CUED/F-INFENG/TR 133, Cambridge, U.K., 1993.

[7] M. Gales and S. J. Young, "Segmental hidden Markov models", *Proc. Eurospeech*, pp. 1579-1582, 1995.

[8] 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.

[9] H. Gish and K. Ng, "A segmental speech model with applications to non-stationary states: an application to speech recognition", *Proc. Int. Conf. Acoust., Speech, Signal Processing*, pp. 447-450, 1993.

[10] J. Goldberger and D. Burshtein, "Scaled random trajectory segmental models", *Computer Speech and Language*, to be published.

[11] W. D. Goldenthal, "Statistical trajectory models for phonetic recognition", Ph.D thesis, MIT, 1994.

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

[13] W. Holmes and M. Russell, "Speech recognition using a linear dynamic segmental HMMs", *Proc. Eurospeech*, pp. 1611-1614, 1995.

[14] W. Holmes and M. Russell, "Modeling speech variability with segmental HMMs", *Proc. Int. Conf. Acoust., Speech, Signal Processing*, pp. 447-450, 1996.

[15] P. Kenny, M. Lennig and P. Mermelstein, "A linear predictive HMM for vector valued observation with application to speech recognition", *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 38, pp. 220-225, 1990.

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

[17] R. G. Leonard, "A database for speaker independent digit recognition", *Proc. Int. Conf. Acoust., Speech, Signal Processing*, pp. 42.11.1-4, Mar. 1984.

[18] M. Ostendorf and S. Roucos, "A stochastic segment model for phoneme-based continuous speech recognition", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1857-1869, 1989.

[19] M. Ostendorf and V. Digalakis, "The stochastic segment model for continuous speech recognition", *Proc. of the 25th Asilomar Conf., Signals, Systems, Computers*, pp. 964-968, 1991.

[20] M. Ostendorf, V. Digalakis and O.A. Kimball, "From HMMs to segmental models: a unified view of stochastic modeling for speech recognition", *IEEE Trans. Speech, Audio Processing*, vol. 38, pp. 360-377, 1996.

[21] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", *Proc. IEEE*, vol. 77, pp. 257-285, 1989.

[22] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.

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

[24] H. Sakoe and S. Chiba, "Dynamic programming optimization for spoken word recognition", *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 26, pp. 43-49, 1978.

[25] R. Schwartz and Y. Chow, "A comparison of several approximate algorithms for finding multiple (N-best) sentence hypotheses", *Proc. Int. Conf. Acoust., Speech, Signal Processing*, pp. 701-704, 1991.

[26] C. J. Wellekens, "Explicit correlation in hidden Markov models for speech recognition", *Proc. Int. Conf. Acoust., Speech, Signal Processing*, pp. 384-387, 1987.

List of Figures

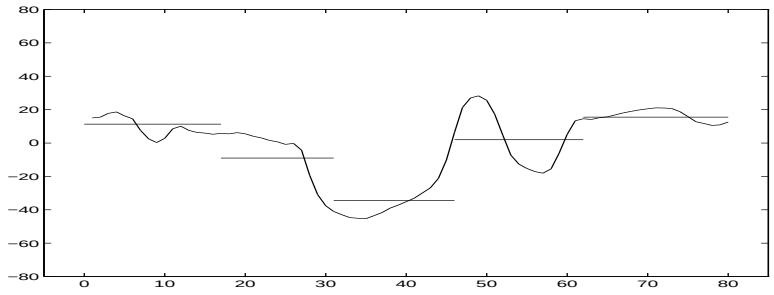

Figure 1: The sixth cepstral coefficient of the acoustic vector and its HMM approximation.

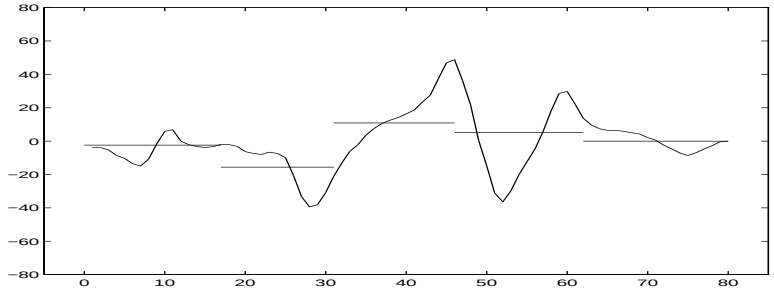

Figure 2: Same as Fig. 1 for the derivative in time of the sixth cepstral coefficient

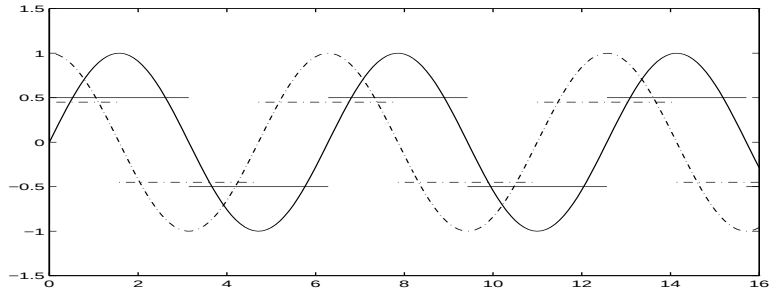

Figure 3: Piecewise constant approximation of a sinusoid and its derivative

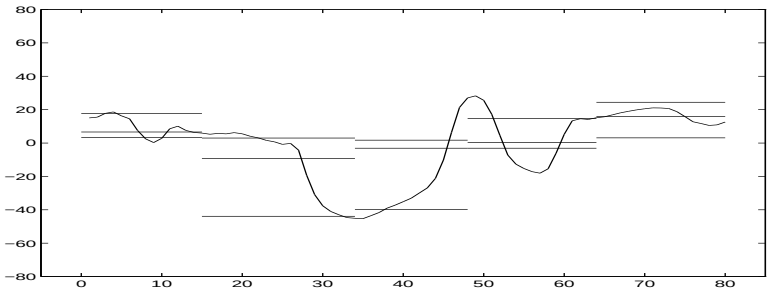

Figure 4: HMM with 5 states and 3 mixtures that models the sixth cepstral coefficient

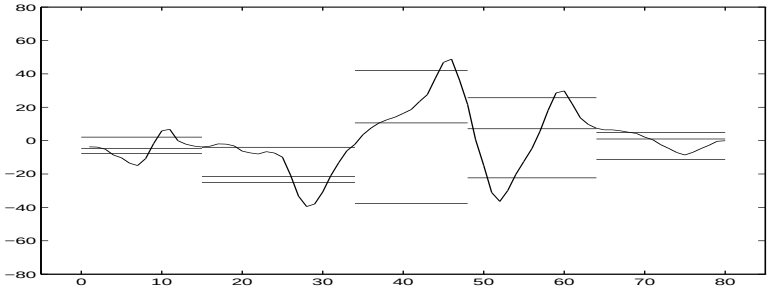

Figure 5: HMM with 5 states and 3 mixtures that models the derivative of the sixth cepstral coefficient

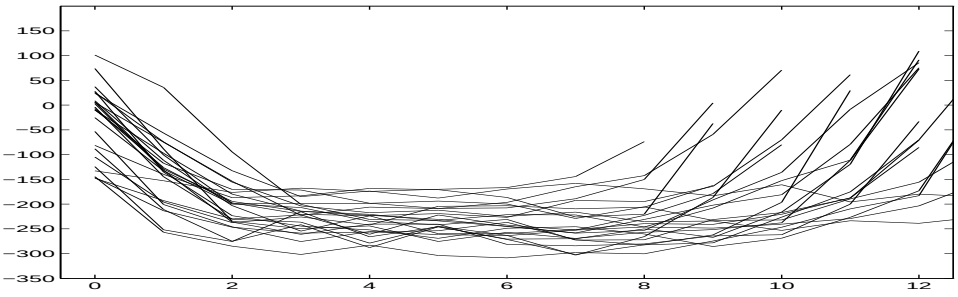

Figure 6: Original data

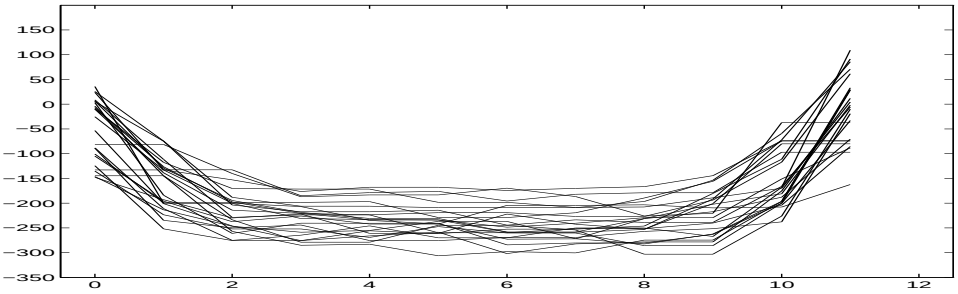

Figure 7: Data after non linear time warping

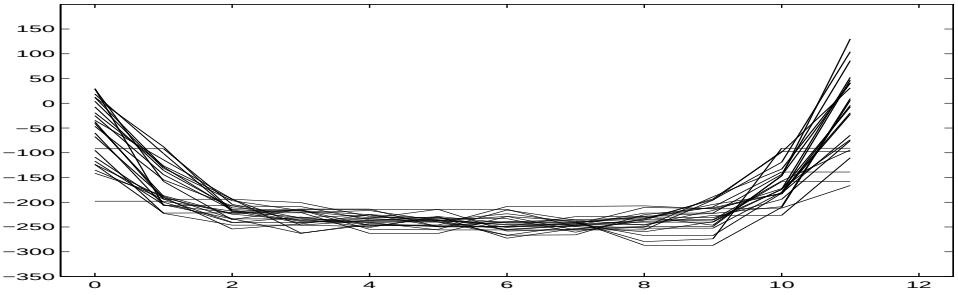

Figure 8: Data after non linear time warping and displacement elimination

List of Tables

	mix1	mix2	mix3
mix1	394	0	0
mix2	202	582	6
mix3	2	216	1540

Table 1: Transition matrix of the three mixtures of the second state presented in Fig. 4

	mix1	mix2	mix3
mix1	196	0	0
mix2	202	212	6
mix3	2	216	212

Table 2: Same as Table 1, except that a sequence of consecutive self transitions is counted only once

model	word error
HMM acoustics	22.1
segmental acoustics	21.4

Table 3: Word error rate results without language model.

model	word error
HMM acoustics + linguistics	8.1
Segmental acoustics + linguistics	11.8
HMM acoustics + segmental acoustics + linguistics	7.8

Table 4: Word error rate results with language model.

	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 5: Triphone recognition rate results

jacob Goldberger received B.Sc. degree in mathematics and computer sciences in 1985 from Bar-Ilan University, M.Sc. degree in mathematics in 1989 and Ph.D. degree in electrical engineering in 1998, both from the Tel-Aviv university, Israel. In 1984-1989, he was a researcher in the Research Laboratories, Israel Ministry of Defense. During 1993-1995 he established and managed the speech recognition department at A.R.T Israel. His research interests include parameter estimation, speech recognition, pattern recognition and statistical data modeling.

David Burshtein (M'92) received the B.Sc. (summa cum laude) and Ph.D. degrees in electrical engineering from Tel-Aviv University, Tel-Aviv, Israel, in 1982 and 1987, respectively

In 1982-1987, he was a Senior Research Engineer in the Research Laboratories, Israel Ministry of Defense, and was involved in research and development of digital signal processing systems. During 1988-1989 he was a Research Staff Member in the Speech Recognition Group of the IBM, T. J. Watson Research Center. In October 1989 he joined the Department of Electrical Engineering – Systems, Tel-Aviv University, where he is currently a faculty member. In the past six years, he has also been acting as a consultant to the Israeli Ministry of Defense and industry. His research interests include parameter estimation, speech processing, speech recognition, statistical pattern recognition and neural networks.

Horacio Franco received the Engineer degree in electronics in 1978 and the Doctor in Engineering degree in 1996, both from the University of Buenos Aires, Argentina. From 1982 to 1988 he was a research fellow at the Laboratory of Sensory Research, Buenos Aires, working on speech synthesis, analysis and recognition. Since 1990 he has been a member of the Speech Technology and Research Laboratory at SRI International, working on acoustic modeling, hybrid Neural Net-HMM speech recognition systems, and in applications of speech technology to language instruction. He has taught introductory courses in Speech Recognition at Stanford, and Neural Networks at San Jose State University. He also lectures regularly at the University of Buenos Aires. He is currently member of the IEEE Neural Networks for Signal Processing comittee.