Modeling Acoustic Transitions in Speech by State-Interpolation Hidden Markov Models

Li Deng, *Senior Member, IEEE*, Patrick Kenny, Matthew Lennig, *Senior Member, IEEE*, and Paul Mermelstein, *Senior Member, IEEE*

***Abstract*—We present a new type of HMM for vowel-to-consonant (VC) and consonant-to-vowel (CV) transitions based on the locus theory of speech perception. The parameters of the model can be trained automatically using the Baum–Welch algorithm and the training procedure does not require that instances of all possible CV and VC pairs be present. When incorporated into an isolated word recognizer with a 75 000 word vocabulary we find that it leads to a modest improvement in recognition rates.**

I. Introduction

THE problem of acoustic variability due to phonetic context is a formidable obstacle to the construction of phoneme-based speech recognizers with very large vocabularies (on the order of 100 000 words or more).

The technique of triphone modeling which has been widely used in recognizers having medium-sized vocabularies (on the order of 1000 words) [3] does not offer a satisfactory solution. In a lexicon containing 86 000 words and 92 000 phonemic transcriptions we counted more than 17 000 triphones. (This figure does not include triphones spanning word boundaries.) In order to train this number of HMM's in a speaker-dependent system, the amount of natural text that a speaker would have to dictate would probably be equivalent to a full-length novel. (Adequate coverage could be obtained in a speaker-independent system, but only at the cost of making the recognition task much more difficult.) Attempts to keep the training set size within reasonable bounds and reduce the number of models by clustering triphones based on prior linguistic knowledge have been found to give only meagre improvements over context-independent models [6], [7]. Automatic clustering algorithms can only be used to cluster those triphones which are represented in the training data and so are not applicable. In any case, the method fails to address the problem of coarticulation between phonemes which are not in immediate juxtaposition.

Manuscript received January 21, 1989; revised March 28, 1990. This work was supported by the Natural Sciences and Engineering Research Council of Canada.

L. Deng was with INRS-Telecommunications, Montreal, Que., Canada H3E 1H6. He is now with the Department of Electrical Engineering, University of Waterloo, Waterloo, Ont., Canada.

P. Kenny, M. Lennig, and P. Mermelstein are with INRS-Telecommunications, Montreal, Que., Canada H3E 1H6.

IEEE Log Number 9104876.

Although the phoneme, by its definition, is a very appropriate unit for large-vocabulary speech recognition the problem of acoustic variability is so serious that some workers advocate abandoning it altogether in favor of a system of units consisting of a large number of small, acoustically stable subword segments [1], [10]. This approach has the drawback of requiring transcriptions in terms of the new units for each word in the lexicon, so that an open-ended lexicon is impossible in principle, and it is incapable of dealing with coarticulation between words in continuous speech.

What is clearly needed is a way of modeling the behavior of phonemes in contexts which are not covered in the training data. In this paper we report the results of an experiment where we attempted to deal with this problem insofar as it applies to the transitional behavior exhibited by vowels in consonant environments.

It is well known that the formant transitions in vowels contain important information for the recognition of consonants, especially stops. A striking example is the case of a word spoken in isolation which ends in a stop. If the stop is not released then, since the stop closure is indistinguishable from the silence which follows the word, all of the information for recognizing the stop is contained in the preceding vowel. If the vowel is modeled by a standard context-independent Markov model then this information is lost.

Our modeling assumption is essentially the locus theory [5] as originally formulated in the 1950's. We assume that for each consonant there is a single target spectrum or "locus" with the property that in VC and CV transitions the vowel spectra tend to converge towards the target for the consonant. The locus has nothing to do with the noise spectrum of the consonant—its existence is inferred from looking at vowel spectra—and it is an empirical fact that, in general, the target is not reached.

In Section II, we show how the locus theory can be used to construct a context-dependent stochastic model for vowels which we call a state-interpolation HMM. Its parameters, including those which represent the consonant loci, can be trained automatically by an extension of the Baum–Welch algorithm. In Section III we give recognition results for the state interpolation HMM and compare them to those obtained by standard context-independent HMM's and generalized triphone models.

II. Representing Acoustic Transitions in Vowels by State-Interpolation HMM's

A. Formulation of the State-Interpolation HMM

The standard HMM's referred to in this paper are context-independent left-to-right hidden Markov models having a unimodal Gaussian distribution associated with each state.

We have found that it is not necessary to train individual covariance matrices for the different states in a standard vowel HMM; we obtain essentially the same recognition performance whether we have one covariance matrix per state, one covariance per vowel or a single covariance for all vowels. (However, in the case of individual covariance matrices for each state we have observed that the variance tends to be relatively small in the middle of the vowel and that it increases monotonically as we move towards either end. This confirms that the standard vowel models are doing a poor job of modeling the data in the transitional regions.) Thus in constructing context-dependent vowel models we focus our attention on the mean vectors.

To explain how we model the transitional regions in vowels, consider the case of a CV transition as illustrated in Fig. 1. (The situation in the case of VC transitions is a mirror image.) For each consonant c we construct a single feature vector ν_c which we call the locus of the consonant. For each vowel v, we designate states $0, \cdots, K-1$ as transitional states and make their mean vectors depend on both c and v. We call state K the vowel steady state; its mean vector μ_v is context independent. The mean vectors at the transitional states, $m_v(k)$, $k = 0, \cdots, K-1$ are obtained by linearly interpolating between the consonant's locus and the vowel's steady state:

$$m_v(k) = \lambda_k \nu_c + \overline{\lambda_k} \mu_v \tag{1}$$

where $\overline{\lambda_k} = 1 - \lambda_k$. The locus vectors themselves do not serve as mean vectors for any of the Gaussian distributions in our word models (which are constructed in the usual way by concatenating standard consonant HMM's with state-independent vowel HMM's). Nevertheless they can be estimated on the same footing as the vowel steady state vectors and the means of all the other output distributions by an extension of the Baum–Welch algorithm for multivariate Gaussian HMM's.

The interpolation weights λ_k cannot be so easily estimated as the standard procedure leads to a set of nonlinear equations, so they were fixed *a priori*. In the experiments reported in Section III we used two transitional states at either end of each vowel model with interpolation weights of $1/3$ and $2/3$.

This implementation has the virtue of simplicity but there are obviously many other possibilities. It might be reasonable to use different numbers of transitional states and different interpolation weights for vowels following aspirated and unaspirated stops. Klatt's work on formant synthesis [9] suggests that it might be better to have several loci for each consonant—perhaps 3, one for plain vowels, one for palatal vowels, and one for rounded vowels. Also, since the linearity assumption implicit in (1) is questionable, it has been suggested [11] that we construct several vectors $\nu_{c,k}$, $k = 0, \cdots, K-1$, for each consonant and assume that

$$m_v(k) = \lambda_k \nu_{c,k} + \overline{\lambda_k} \mu_v.$$

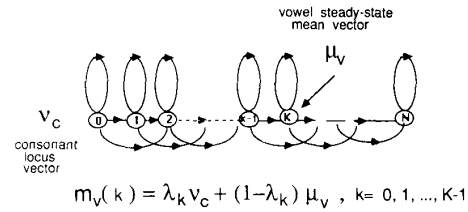


Fig. 1. A vowel HMM having an acoustic transition structure. The leftmost K states (labeled 0 to $K-1$) represent the consonant-to-vowel transition. The state labeled K (shown) and a few states to the right of it (omitted for clarity) represent the vowel's stationary portion. The remaining states represent the vowel-to-consonant transition (omitted also).

In presenting the reestimation formulas we consider only the case of CV transitions and, for the sake of simplicity, we assume that there is only one token for every CV pair in the training data. The reader may find it useful to work out the derivation in the case where the training data consists of a single CV pair. This results in a pair of simultaneous equations involving the consonant locus and the vowel steady state. In the general case (presented in Section II-B), the reestimation procedure leads to a large system of simultaneous equations involving all of the consonant loci and all of the vowel steady states but, since it's a linear system, there is no difficulty in solving it.

B. The Objective Function for Maximization

For a Gaussian HMM [12], it can be easily shown that maximizing the following objective function will guarantee an increase in the likelihood of the observations:

$$Q = \sum_{v=1}^{V} \sum_{c=1}^{C} \sum_{t=1}^{T} \sum_{s=1}^{N} \gamma_t^{cv}(s) \ln D(Y_t^{cv}, s) + \sum_{c=1}^{C} \sum_{v=1}^{V} \sum_{t=1}^{T} \sum_{s,s'} \gamma_t^{cv}(s, s') \ln P(s'|s) \tag{2}$$

where V and C are the total number of vowels and consonants, respectively; $P(s'|s)$ is the state transition probabilities from state s to state s'; $D(Y_t^{cv}, s)$ is a d-dimensional multivariate Gaussian distribution associated with state s for Y_t^{cv}, the observation vector at time t in a CV context; $\gamma_t^{cv}(s, s')$ and $\gamma_t^{cv}(s)$ are the conditional probabilities that a state transition from s to s' takes place at time t, and that state s is occupied at time t, respectively, given that an observation sequence in a cv context is generated by the model. These conditional probabilities can be computed efficiently from the forward and backward probabilities in a standard way [2], [12].

In maximizing the Q function in (2), the two terms can be treated separately. The second term involves only the

transition probabilities, and the reestimation formula for them is the same as that for the standard HMM and will be omitted here.

By ignoring $\ln (2\pi)^{d/2}$ in the multivariate Gaussian distribution function and restricting to only the $(K+1)$ leftmost states in the HMM's for vowels (as shown in Fig. 1), we rewrite the objective function as

$$Q' = \sum_{c=1}^{C} \sum_{v=1}^{V} \sum_{t=1}^{T} \sum_{k=0}^{K} \gamma_t^{cv}(k) \{-\tfrac{1}{2} \ln |\Sigma_v(k)| - \tfrac{1}{2} (Y_t - m_v(k))^* \Sigma_v^{-1}(k) (Y_t - m_v(k))\} \quad (3)$$

where $\Sigma_v(k)$ is the covariance matrix associated with state k in the state-interpolation HMM for vowel v. Substituting the interpolation relation in (1) into (3), we have

$$\begin{aligned} Q' = & \sum_{c,v} \sum_{t} \gamma_t^{cv}(K) \{-\tfrac{1}{2} \ln |\Sigma_v(K)| \\ & - \tfrac{1}{2} (Y_t - \mu_v)^* \Sigma_v^{-1}(K) (Y_t - \mu_v)\} \\ & + \sum_{c,v} \sum_{t} \sum_{k=0}^{K-1} \gamma_t^{cv}(k) \{-\tfrac{1}{2} \ln |\Sigma_v(k)| \\ & - \tfrac{1}{2} (Y_t - \lambda_k \nu_c - \bar{\lambda}_k \mu_v)^* \\ & \cdot \sum_{v} (k)^{-1} (Y_t - \lambda_k \nu_c - \bar{\lambda}_k \mu_v)\}. \end{aligned} \quad (4)$$

C. Reestimation of Consonant Locus Vectors and Vowel Steady-State Mean Vectors

At a critical point, the derivatives of the Q' function in (4) with respect to ν_c and μ_v must be all zero.

From

$$\frac{\partial Q'}{\partial \nu_c} = 0$$

we have

$$\sum_{v} \sum_{t} \sum_{k=0}^{K-1} \gamma_t^{cv}(k) \Sigma_v^{-1}(k) (Y_t - \lambda_k \nu_c - \bar{\lambda}_k \mu_v) \lambda_k = 0$$
$$c = 1, 2, \cdots, C. \quad (5)$$

From

$$\frac{\partial Q'}{\partial \mu_v} = 0$$

we have

$$\sum_{c} \sum_{t} \sum_{k=0}^{K-1} \gamma_t^{cv}(k) \Sigma_v^{-1}(k) (Y_t - \lambda_k \nu_c - \bar{\lambda}_k \mu_v) \bar{\lambda}_k = 0$$
$$v = 1, 2, \cdots, V. \quad (6)$$

To achieve a tractable solution to (5) and (6), assume that the covariance matrices Σ's are the same for all states and for all vowels. This assumption converts (5) and (6) to a symmetric linear system of equations, which can be expressed in the following matrix form:

$$\begin{bmatrix} A_1 & & & B_{11} & B_{12} & \cdots & B_{1V} \\ & A_2 & O & B_{21} & B_{22} & \cdots & B_{2V} \\ & \cdots & & \cdots & \cdots & \cdots & \cdots \\ & & A_C & B_{C1} & B_{C2} & \cdots & B_{CV} \\ & & & E_1 & & & \\ & * & & & E_2 & & O \\ & & & & & \cdots & \\ & & & & & & E_V \end{bmatrix} \cdot \begin{bmatrix} \nu_1 \\ \nu_2 \\ \cdots \\ \nu_C \\ \mu_1 \\ \mu_2 \\ \cdots \\ \mu_V \end{bmatrix} = \begin{bmatrix} U_1 \\ U_2 \\ \cdots \\ U_C \\ W_1 \\ W_2 \\ \cdots \\ W_V \end{bmatrix} \quad (7)$$

where

$$U_c = \sum_{v} \sum_{t} \sum_{k=0}^{K-1} \lambda_k \gamma_t^{cv}(k) Y_t, \qquad c = 1, 2, \cdots, C$$

$$W_v = \sum_{c} \sum_{t} \sum_{k=0}^{K-1} \bar{\lambda}_k \gamma_t^{cv}(k) Y_t + \sum_{c} \sum_{t} \gamma_t^{cv}(K) Y_t$$
$$v = 1, 2, \cdots, V$$

$$A_c = \sum_{v} \sum_{t} \sum_{k=0}^{K-1} \lambda_k^2 \gamma_t^{cv}(k), \qquad c = 1, 2, \cdots, C$$

$$E_v = \sum_{c} \sum_{t} \sum_{k=0}^{K-1} \bar{\lambda}_k^2 \gamma_t^{cv}(k) + \sum_{c} \sum_{t} \gamma_t^{cv}(K)$$
$$v = 1, 2, \cdots, V$$

$$B_{cv} = \sum_{t} \sum_{k=0}^{K-1} \lambda_k \bar{\lambda}_k \gamma_t^{cv}(k) \qquad v = 1, 2, \cdots, V$$
$$c = 1, 2, \cdots, C.$$

The solution to this linear system gives the Baum–Welch reestimates for the locus vectors ν_c, $c = 1, 2, \cdots, C$ and the vowel steady-state mean vectors μ_v, $v = 1, 2, \cdots, V$.

D. Reestimation of the Covariance Matrix

Once the locus vector for each consonant and the steady-state mean vector for each vowel are reestimated according to (7), the reestimation formula for the covariance matrix for vowels (pooled across all states and

across all vowels), Σ, can be easily obtained. First let $T = \Sigma^{-1}$ and regard the objective function Q' in (2) as a function of T. It is well known that the derivative of $\ln |T|$ with respect to its ijth entry, t_{ij}, is the ijth entry of Σ, σ_{ij} [12]. To find a critical point, setting

$$\frac{\partial Q'}{\partial \sigma_{ij}} = 0$$

and writing the result in a matrix form, we have

$$\Sigma = \left(\sum_c \sum_t \sum_{k=0}^{K-1} \gamma_t^{cv}(k)(Y_t - \lambda_k \nu_c - \overline{\lambda}_k \mu_v) \cdot (Y_t - \lambda_k \nu_c - \overline{\lambda}_k \mu_v)^* + \sum_c \sum_t \gamma_t^{cv}(K)(Y_t - \mu_v)(Y_t - \mu_v)^* \right) \Bigg/ \sum_c \sum_t \sum_{k=0}^{K} \gamma_t^{cv}(k). \quad (8)$$

E. Extensions

Equations (7) and (8) for reestimating the state-interpolation HMM parameters are applicable only to the vowel tokens in a CV context. For VC contexts, similar reestimation equations can be obtained for the mean vectors of the K rightmost states in the vowel state-interpolation HMM, if we assume two independent consonant locus vectors, one preceding and one following the vowel. When a vowel token occurs in a neutral context, that is, in a word boundary, or is adjacent to another vowel or to /h/,[1] then the assumption of interpolation with a locus vector is invalid. In this case, we assume that the mean vectors of the K leftmost or the K rightmost states in the vowel state-interpolation HMM are the same as the mean vectors corresponding to the HMM states representing the vowel steady state. This treatment of the neutral context has somewhat complicated the reestimation formulas, which are described in the Appendix.

III. Speech Recognition Experiments

We have implemented the state-interpolation HMM's, described in Section II and the Appendix, in an isolated-word speaker-trained 75 000-word recognizer. The recognition algorithm consists of word endpoint detection (with manual adjustment where necessary), a fast search to generate a list of most likely word choices, and computation of exact likelihoods for these choices [8]. State-interpolation HMM's are used only at the exact likelihood scoring stage, since complete phonetic transcriptions are needed to fill in the mean vectors associated with vowel transitional states. In the experiments reported here, no language model was used (so homophone confusions are not considered as errors).

[1]For our purposes it is inappropriate to treat /h/ as a consonant. The locus theory is clearly not applicable to /h/ since it is typically realized as a voiceless version of the following vowel and so does not have a well defined place of articulation.

As mentioned in Section II, we use two transitional states at either end of each vowel model with interpolation weights of $\frac{1}{3}$ and $\frac{2}{3}$. (An exception is made for /ə/ which is modeled using a 4-state standard HMM rather than a state-interpolation HMM.) Table I reports the results of recognition experiments on 5 speakers.

The acoustic parameters used are mel-based cepstral coefficients and their differences, referred to below as C's and ΔC's, respectively, as well as the difference in the loudness $\Delta C0$ [4]. These are extracted every 10 ms using a Hamming window with a width of 25.6 ms. Training and test data consist of short extracts of novels, newspaper articles, etc.

Although the state-interpolation model performs at least as well as the standard model in every case, the improvement is not very big. This is only to be expected as the two models differ only in the way they handle VC and CV transitions (which is to say less than 10% of the frames).

The linear interpolation assumption implicit in (1) is not strictly satisfied by cepstral coefficients (particularly those of high order), so we decided to experiment with other parameter sets. Table II shows recognition reuslts on one of the speakers using a variety of different acoustic parameters including log area ratios obtained from a standard LPC analysis (denoted by LAR's), and centers of gravity for frequencies within critical bands (denoted by FREQ's). The interpolation model performs slightly better than the standard model in almost all cases but the ranking of the parameter sets with the state-interpolation HMM's follows that of the standard HMM's.

As explained in Section II-C, it is convenient to have a single covariance matrix for all states in all vowel models. We know that the standard model produces a high variance for vowel transitional states when individual covariance matrices are trained for each state, although this does not have a material effect on recognition results. To investigate the role of the covariance matrix in the state-interpolation model, we tried an experiment on one speaker where we trained two full covariance matrices for the vowels, one for the transitional states and one for the states whose means are context independent. Comparing the first and second lines of Table III shows that this leads to a degradation in performance for both the standard and state-interpolation HMM's.

The second line of Table II shows that when the parameter set is limited to static cepstral coefficients, the improvement obtained by the state-interpolation model is relatively large. This is not surprising since our modeling assumptions are clearly much more appropriate for the static cepstral coefficients than for their differences. It suggests that it might be best to use the interpolation model for the static coefficients and an ordinary HMM for the dynamic coefficients. In order to implement this we found it necessary to decorrelate the static and dynamic coefficients, i.e., to assume that the covariance matrix is block diagonal. The result shown in the third line of Table III demonstrated that when implemented this way, the

TABLE I
COMPARISON OF RECOGNITION RATES FOR 5 SPEAKERS

Speaker	Training Data (No. of Words)	Test Data (No. of Words)	Standard HMM	Interpolation HMM
ML (male)	1203	782	84.5%	86.4%
AM (male)	1100	483	54.4%	59.6%
AM	2039	483	68.2%	68.2%
AM	2742	483	68.7%	68.6%
CA (female)	717	1090	67.9%	71.5%
CA	1532	1090	70.2%	71.2%
AMM (female)	1600	586	79.0%	79.8%
MG (female)	1194	588	75.3%	76.7%

TABLE II
COMPARISON OF ROCOGNITION RATES FOR THE TWO TYPES OF HMM USING VARIOUS SETS OF ACOUSTIC PARAMETERS. SPEAKER ML

Parameter Sets	Standard HMM's	State-Interpolation HMM's
7 C's + 7 ΔC's + $\Delta C0$	84.5%	86.4%
7 C's	70.0%	74.0%
7 C's + $\Delta C0$	79.0%	81.0%
6 LAR's	52.8%	55.1%
12 LAR's	67.1%	65.5%
12 LAR's + $\Delta C0$ + 5 ΔLAR's	70.1%	70.0%
7 C's + 8 ΔC's + 3 LAR's	81.1%	84.0%
7 FREQ's	50.0%	55.0%
7 FREQ's + 7 ΔFREQ's + $\Delta C0$	65.9%	69.0%

TABLE III
COMPARISON OF RECOGNITION RATES USING VARIOUS COVARIANCE MATRICES. SPEAKER ML

Choice of Covariance Matrix	Standard HMM's	State-Interpolation HMM's
One full covariance matrix	84.5%	86.4%
Two full covariance matrices	82.5%	83.8%
Two block diagonal covariance matrices	80.1%	83.4%

TABLE IV
COMPARISON WITH RECOGNITION RESULTS OBTAINED USING GENERALIZED DIPHONE AND TRIPHONE MODELS. SPEAKER AM

Training Data (No. of Words)	Standard HMM	Interpolation HMM	Generalized Diphone HMM's	Generalized Triphone HMM's
1100	54.4%	59.6%	54.0%	53.4%
2039	68.2%	68.2%	73.0%	72.0%
2742	68.7%	68.6%	74.2%	76.1%

state-interpolation model again outperforms the standard model, but the result is not as good as the first implementation.

Our work on context-dependent modeling using (generalized) diphones and triphones has been reported in detail elsewhere [6]. Briefly, we define allophones of each phoneme using a broad classification of its right and left neighbors. (Triphones take both into account, diphones only one.) Table IV shows that, unlike the state-interpolation model, this method only gives improvements with relatively large training sets. (It is also necessary to bear in mind that the interpolation model only attempts to deal with context-dependent variability exhibited by vowels, whereas diphone and triphone models are used for both consonants and vowels.)

IV. CONCLUSION

In this paper, we have described our development of the state-interpolation HMM, guided by the locus theory of speech perception, for context-dependent phonetic modeling. We have found that, when compared to the standard HMM, it leads to modest but consistent improvements in recognition performance, both across speakers and across parameter sets. We have also observed that the pooled variance for vowels in the interpolation model is consistently smaller than in the standard model, indicating a better fit to the training data in the VC and CV transitional regions (the only part of the data which is handled differently by the two models).

These improvements are not as big as might be expected given the universally recognized importance of formant transitions as a cue for the place of articulation of consonants. Several possible explanations suggest themselves. It may be that our acoustic parameters are insufficiently sensitive to formant transitions. It may be that our implementation was too rigid: recall that as mentioned in Section II-A, we did not attempt to deal with the problem of optimizing the interpolation weights. Nonetheless, even a poor stochastic model, provided it is trained automatically, can give good recognition results—think of the standard HMM—since it incorporates a mechanism for dealing with its own inadequacies ("random variation"). It may be that the locus theory will have to be given a more subtle formulation before it can serve as a basis for a good stochastic model. The evidence for the locus theory is primarily from speech perception and formant synthesis experiments; it gives one way of synthesizing acceptable VC and CV transitions, but it does not claim to be a model for speech production.

Whatever the explanation, it is clear that if the problem of speaker-dependent very large vocabulary recognition is to be solved by statistical techniques then it will be necessary to construct robust models capable of generalizing

from small training sets in which we cannot possibly hope to have all types of context-dependent acoustic-phonetic behavior adequately represented. The state-interpolation model is a first step in this direction.

Appendix
Reestimation of Consonant Locus Vectors and Vowel Steady-State Means in Both CV and Neutral Contexts

In Section II, we derive the Baum–Welch reestimation formulas for consonant locus vectors and vowel steady-state means assuming that each vowel token is always in a *cv* context. In this Appendix, we extend the derivation to include the realistic situation where a vowel token can be in either a *cv* context, or in a neutral context. In the case of the *cv* context, the Gaussian mean vectors associated with the HMM states intended to model acoustic transitions are determined by (1). For the neutral context, these mean vectors are assumed to be the same as the mean vector associated with an HMM state representing the vowel steady state.

Due to the presence of neutral contexts, the objective function Q in (4) will have additional terms corresponding to the neutral contexts. The new Q function is

$$\begin{aligned} Q = & \sum_c \sum_v \sum_t \gamma_t^{cv}(K) \{ -\tfrac{1}{2} \ln |\Sigma_v(K)| \\ & - \tfrac{1}{2} (Y_t - \mu_v)^* \Sigma_v^{-1}(K) (Y_t - \mu_v) \} \\ & + \sum_c \sum_v \sum_t \sum_{k=0}^{K-1} \gamma_t^{cv}(k) \{ -\tfrac{1}{2} \ln |\Sigma_v(k)| \\ & - \tfrac{1}{2} (Y_t - \lambda_k \nu_c - \overline{\lambda}_k \mu_v)^* \\ & \cdot \Sigma_v^{-1}(k) (Y_t - \lambda_k \nu_c - \overline{\lambda}_k \mu_v) \} \\ & + \sum_n \sum_v \sum_t \sum_{k=0}^{K} \gamma_t^{nv}(k) \left\{ -\tfrac{1}{2} \ln |\Sigma_v(k)| \right. \\ & \left. - \tfrac{1}{2} (Y_t - \mu_v)^* \Sigma_v^{-1}(k) (Y_t - \mu_v) \right\} \end{aligned}$$

where the summation over n denotes inclusion of all neutral contexts for vowel tokens, and $\gamma_t^{nv}(k)$ is the conditional probability that the kth state is occupied at time t, given that an observation sequence corresponding to a token of vowel v in the neutral context n is generated by the model. Other notations are identical to those used in (4).

Maximization of the above Q function leads to the following symmetric linear system of equations, whose solution is the Baum–Welch reestimate of the locus vectors, ν_c, $c = 1, 2, \cdots, C$ and the vowel steady-state mean vectors μ_v, $v = 1, 2, \cdots, V$:

$$\begin{bmatrix} A_1 & & & B_{11} & B_{12} & \cdots & B_{1V} \\ & A_2 & O & B_{21} & B_{22} & \cdots & B_{2V} \\ & \cdots & & \cdots & \cdots & \cdots & \cdots \\ & & A_C & B_{C1} & B_{C2} & \cdots & B_{CV} \\ & & & E_1 & & & \\ * & & & & E_2 & & O \\ & & & & & \cdots & \\ & & & & & & E_V \end{bmatrix} \begin{bmatrix} \nu_1 \\ \nu_2 \\ \cdots \\ \nu_C \\ \mu_1 \\ \mu_2 \\ \cdots \\ \mu_V \end{bmatrix} = \begin{bmatrix} U_1 \\ U_2 \\ \cdots \\ U_C \\ W_1 \\ W_2 \\ \cdots \\ W_V \end{bmatrix}$$

where

$$U_c = \sum_v \sum_t \sum_{k=0}^{K-1} \lambda_k \gamma_t^{cv}(k) Y_t, \qquad c = 1, 2, \cdots, C$$

$$\begin{aligned} W_v = & \sum_c \sum_t \sum_{k=0}^{K-1} \overline{\lambda}_k \gamma_t^{cv}(k) Y_t + \sum_c \sum_t \gamma_t^{cv}(K) Y_t \\ & + \sum_n \sum_t \sum_{k=0}^{K} \gamma_t^{nv}(k) Y_t, \qquad v = 1, 2, \cdots, V \end{aligned}$$

$$A_c = \sum_v \sum_t \sum_{k=0}^{K-1} \lambda_k^2 \gamma_t^{cv}(k), \qquad c = 1, 2, \cdots, C$$

$$\begin{aligned} E_v = & \sum_c \sum_t \sum_{k=0}^{K-1} \overline{\lambda}_k^2 \gamma_t^{cv}(k) + \sum_c \sum_t \gamma_t^{cv}(K) \\ & + \sum_n \sum_t \sum_{k=0}^{K} \gamma_t^{nv}(k) Y_t, \qquad v = 1, 2, \cdots, V \end{aligned}$$

$$B_{cv} = \sum_t \sum_{k=0}^{K-1} \lambda_k \overline{\lambda}_k \gamma_t^{cv}(k), \qquad v = 1, 2, \cdots, V$$

$$c = 1, 2, \cdots, C$$

References

[1] L. R. Bahl, P. F. Brown, P. V. de Souza, R. L. Mercer, and M. A. Picheny, "Acoustic Markov models used in the Tangora speech recognition system," in *Proc. IEEE ICASSP*, vol. 1, 1988, pp. 497–500.

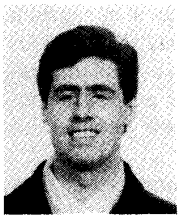
[2] L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," *Inequalities*, vol. 3, pp. 1–8, 1972.

[3] Y. L. Chow, M. D. Dunham, O. A. Kimball, M. A. Krasner, G. F. Kubala, J. Makhoul, P. J. Price, S. Roucos, and R. M. Schwartz,

"BYBLOS: The BBN continuous speech recognition system," in *Proc. IEEE ICASSP*, vol. 1, 1987, pp. 89–91.
[4] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, no. 4, pp. 357–365, 1980.
[5] P. Delattre, A. M. Liberman, and F. S. Cooper, "Acoustic loci and transitional cues for consonants," *J. Acoust. Soc. Amer.*, vol. 27, pp. 769–774, 1955.
[6] L. Deng, M. Lennig, F. Seitz, and P. Mermelstein, "Large vocabulary word recognition using context-dependent allophonic hidden Markov models," *Computer Speech Language*, vol. 4, no. 4, pp. 345–357, Dec. 1990.
[7] A. Derouault, "Context-dependent phonetic Markov models for large vocabulary speech recognition," in *Proc. IEEE ICASSP*, vol. 1, 1987, pp. 360–363.
[8] V. Gupta, M. Lennig, and P. Mermelstein, "Fast search strategy in a large vocabulary word recognizer," *J. Acoust. Soc. Amer.*, vol. 84, pp. 2007–2017, 1988.
[9] D. H. Klatt, "Review of text-to-speech conversion for English," *J. Acoust. Soc. Amer.*, vol. 82, pp. 737–793, 1987.
[10] C. H. Lee, F. K. Soong, and B. H. Juang, "A segment model approach to speech recognition," in *Proc. IEEE ICASSP*, vol. 1, 1988, pp. 501–504.
[11] G. Doddington, personal communication.
[12] L. A. Liporace, "Maximum likelihood estimation for multivariate observations of Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 729–734, 1982.
[13] D. B. Paul and E. A. Martin, "Speaker stress-resistant continuous speech recognition," in *Proc. IEEE ICASSP*, vol. 1, pp. 283–286, 1988.

Li Deng (S'83–M'86–SM'91) received the Ph.D. degree in electrical engineering from the University of Wisconsin-Madison in 1986.

He worked on large vocabulary speech recognition at INRS-Telecommunications, Montreal, Que., Canada, from 1986 to 1989. Since 1989, he has been with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ont., Canada, as an Assistant Professor. His research interests include acoustic-phonetic modeling of speech, statistical signal modeling, auditory signal processing, and auditory neuroscience. In these areas, he has written over 40 published papers.



Patrick Kenny was born in Montreal, Canada, in 1955. He graduated from Trinity College, Dublin, Ireland, with first class honors in mathematics in 1976 and received the M.Sc. and Ph.D. degrees, also in mathematics, from McGill University.

He has been working in speech recognition at INRS-Télécommunications, Université due Québec, since 1986, and he is particularly interested in stochastic modeling of the speech signal.

Matthew Lennig (M'79–SM'85) received the A.B. degree (summa cum laude) from Princeton University in 1974 in theoretical linguistics; he received the Ph.D. degree in sociolinguistics in 1978 from the University of Pennsylvania, which he attended as a National Science Foundation Fellow; and the M.Eng. degree in electrical engineering from McGill University in 1984.

He joined Bell-Northern Research in 1978 and is currently the Manager of Interactive Voice Systems, with responsibility for the development of the interactive voice technology component of Northern Telecom's AABS product, which uses speech recognition in the telephone network to automate collect and third-number billed calls. Prior to 1988 he was the Manager of Speech Systems, with responsibility for development of Bell Canada's speech-recognition-based 976 Directory and for algorithmic research in the areas of speech recognition, speaker verification, and speech synthesis. Since 1981 he has been a Visiting Professor at l'Institut National de la Recherche Scientifique en Télécommunications (Université du Québec) where he currently directs a research project on very large (86 000-word) vocabulary speech recognition.

Paul Mermelstein (S'58–M'63–SM'77) was born in Czechoslovakia in 1939. He received the B.Eng. degree in engineering physics from McGill University, Montreal, Que., Canada, in 1959, and the S.M., E.E., and D.Sc. degrees from the Massachusetts Institute of Technology, Cambridge, in 1960, 1963, and 1964, respectively.

From 1964 to 1973 he was a member of the Technical Staff in the Speech and Communications Research Department of Bell Laboratories, Murray Hill, NJ. From 1973 to 1977 he was a member of the Research Staff of Haskins Laboratories, conducting research in speech analysis, perception, and recognition. Over the years 1977 to 1986 he was Manager of Speech Communication Systems at Bell-Northern Research in Montreal. He currently serves as Manager of Man-Machine Systems at BNR. He served as Associate Editor for Speech Processing for the *Journal of the Acoustical Society of America* from 1983 to 1987. He has also served on numerous expert groups on speech coding of CCITT WP8 recommending standards on speech coding for telecommunication applications. Since 1977 he has held appointments as Visiting Professor at INRS-Télécommunications (Université du Québec) and Auxiliary Professor of Electrical Engineering at McGill University.

Dr. Mermelstein is Editor for Speech Communication of the IEEE TRANSACTIONS ON COMMUNICATIONS.