GAUSSIAN MIXTURE MODELS OF PHONETIC BOUNDARIES FOR SPEECH RECOGNITION

M. Kamal Omar, Mark Hasegawa-Johnson, and Stephen Levinson

University of Illinois at Urbana-Champaign,
Department of Electrical And Computer Engineering,
Urbana, IL 61801.

ABSTRACT

A new approach to represent temporal correlation in an automatic speech recognition system is described. It introduces an acoustic feature set that captures the dynamics of speech signal at the phoneme boundaries in combination with the traditional acoustic feature set representing the periods that are assumed to be quasi-stationary of speech. This newly introduced feature set represents an observed random vector associated with the state transition in HMM. For the same complexity and number of parameters, this approach improves the phoneme recognition accuracy by 3.5% compared to the context-independent HMM models. Stop consonant recognition accuracy is increased by 40%.

1. INTRODUCTION

Two of the main drawbacks of the current HMM model for automatic speech recognition are the assumptions of conditionally independent observations given the state sequence which results in not utilizing the information related to temporal correlation, and the geometric duration model which is far from being an accurate model of the duration of the subunits of speech [1, 2, 3]. The first problem has been addressed by several approaches including using context-dependent models of speech units, implicit trajectory models, segmental acoustic features, explicitly dynamic acoustic features, and using acoustic features related to the temporal correlation within the observation vector associated with the quasi-stationary subunits represented by HMM states, [1]. Other researchers, [2], went further and advocated the use of relatively large temporal segments of speech signal, believing that the information about a phoneme is not localized to the period of that phoneme only. Solutions of the second problem were first introduced in [3] by adding explicit model of the phoneme duration to the HMM model. This model uses Gamma probability density function to model the duration. Other proposed solutions to this problem include a higher-order HMM that gives better model of the duration than the first-order HMM and reduces the effect of the assumed Markovian property of the state transitions, [4].

As HMM has become the dominant model for speech recognition, many researchers, [4], have noticed that the state transition probabilities have a negligible impact on the recognition rate and are often ignored. These observations and the previously mentioned need for a more accurate model of the duration of the phonemes than the geometric model motivate the consideration of the actual rule that should be assigned to the transition probabilities in speech recognition tasks.

One of the main motivations of the recent search for an alternative model for speech recognition is that several experiments proved that human perception of speech is based mainly on certain landmarks in the speech signal [5]. These landmarks identify times when the acoustic manifestations of the linguistically motivated distinctive features are most salient. Most of these landmarks are at the boundaries of the quasi-stationary subunits of speech. This proves that the importance of having a good representation of the probability of transition is at least as important as having a good representation of these quasi-stationary subunits of speech. Studies of acoustic landmarks suggest a new solution to the old problem of HMM transition modeling: perhaps phone transition probabilities in an HMM should be observation-dependent.

This paper presents a representation of these transition probabilities by a mixture of Gaussian probability density functions. They model the probability density function of acoustic measures that are associated with the salient properties of the spectrum at this specific transition. This approach not only makes use of the information in the spectrum at these transitions but also allows an efficient employment of the parameters of the HMM model in the utterance decoding process. The effect of the transition probabilities is no longer negligible compared to the high dimensionality of the observation probability density functions. This approach is flexible in that it allows employing different acoustic features to model different state transitions. The selection of each specific acoustic feature set can be based on phoneme classification research or by using information-theoretic measures for selection.

2. NOMENCLATURE

Throughout this representation, the notation used in [6] will be adopted. Each phoneme in the phoneme set is represented by a three-state left-to-right HMM. Let $A = [a_{ij}]_{n\times n}$ be the state transition matrix. Associated with each state j of the hidden Markov chain is a probability density $b_j(X)$ of the observed d-dimensional random vector X, and with each transition from i to j a probability density $a_{ij}(Y)$ of the l-dimensional random vector Y. The probability densities of both of them are approximated by a mixture of Gaussian probability density functions.

$$b_j(X) = \sum_{k=1}^{m} c_{jk} N(X, \mu_{jk}, U_{jk})$$

where m is known; $c_{jk} \geq 0$ for $1 \leq j \leq n, 1 \leq k \leq m$;

$$\sum_{k=1}^{m} c_{jk} = 1$$

for $1 \leq j \leq n$; and $N(X, \mu, U)$ denote the d-dimensional normal density function of mean vector μ and covariance matrix U.

$$\begin{aligned} a_{ij}(Y) &= Pr(q_t = j|q_{t-1} = i) Pr(Y|q_t = j, q_{t-1} = i) \\ &= Pr(q_t = j|q_{t-1} = i) \sum_{r=1}^{p} w_{ijr} N(Y, \rho_{ijr}, V_{ijr}) \end{aligned}$$

where p is known; $w_{ijr} \geq 0$ for $1 \leq i \leq n, 1 \leq j \leq n, 1 \leq r \leq p$;

$$\sum_{r=1}^{p} w_{ijr} = 1$$

for $1 \leq i \leq n, 1 \leq j \leq n$; and $N(Y, \rho, V)$ denote the l-dimensional normal density function of mean vector ρ and covariance matrix V. Let $O = (O_1, O_2, ..., O_T)$ be a given observation sequence of the vector X and and let $Z = (Z_1, Z_2, Z_3.....Z_T)$ be another given observation sequence of the vector Y. Then given both O and Z and a particular choice of parameter values λ of both Gaussian mixtures, we can efficiently evaluate the likelihood function, $L_\lambda(O, Z)$, by the forward-backward method of Baum. The forward-backward procedure is used to calculate:

$$\alpha_t(i) = Pr(O_1, O_2, ..., O_t, Z_1, Z_2, ..., Z_t, q_t = q_i|\lambda)$$

and

$$\beta_t(i) = Pr(O_{t+1}, O_{t+2}, ...O_T, Z_{t+1}, Z_{t+2}, ...Z_T|q_t = q_i, \lambda)$$

The likelihood function can be written as

$$L_\lambda(O, Z) = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_t(i) a_{ij}(Z_{t+1}) b_j(O_{t+1}) \beta_{t+1}(j)$$

for any t between 1 and T-1.

3. THE ESTIMATION ALGORITHM

The reestimation equations for the HMM parameters are based on [6]. We extended them to calculate the values of the parameters of the Gaussian mixtures representing the state transition probabilities as following.

$$\acute{w}_{ijr} = \frac{\sum_{t=2}^{T} \gamma_t(i,j,r)}{\sum_{t=2}^{T} \sum_{r=1}^{p} \gamma_t(i,j,r)} \quad (1)$$

$$\acute{\rho}_{ijr} = \frac{\sum_{t=2}^{T} \gamma_t(i,j,r) Z_t}{\sum_{t=2}^{T} \gamma_t(i,j,r)} \quad (2)$$

$$\acute{V}_{ijr} = \frac{\sum_{t=2}^{T} \gamma_t(i,j,r)(Z_t - \rho_{ijr})(Z_t - \rho_{ijr})^T}{\sum_{t=2}^{T} \gamma_t(i,j,r)} \quad (3)$$

$$\acute{c}_{jk} = \frac{\sum_{t=1}^{T} \zeta_t(j,k)}{\sum_{t=1}^{T} \sum_{k=1}^{m} \zeta_t(j,k)} \quad (4)$$

$$\acute{\mu}_{jk} = \frac{\sum_{t=1}^{T} \zeta_t(j,k) O_t}{\sum_{t=1}^{T} \zeta_t(j,k)} \quad (5)$$

$$\acute{U}_{jk} = \frac{\sum_{t=1}^{T} \zeta_t(j,k)(O_t - \mu_{jk})(O_t - \mu_{jk})^T}{\sum_{t=1}^{T} \zeta_t(j,k)} \quad (6)$$

where

$$\gamma_t(i,j,r) = \alpha_{t-1}(i) b_j(O_t) w_{ijr} \frac{\partial a_{ij}}{\partial w_{ijr}}|_{z_t} \beta_t(j) \quad (7)$$

for $1 < t \leq T$

$$\zeta_t(j,k) = \begin{cases} c_{jk} \frac{\partial b_j}{\partial c_{jk}}|_{o_t} \beta_t(j), & \text{for t=1} \\ \sum_{i=1}^{n} \alpha_{t-1}(i) a_{ij}(Z_t) c_{jk} \frac{\partial b_j}{\partial c_{jk}}|_{o_t} \beta_t(j) & \text{for } 1 < t \leq T \end{cases} \quad (8)$$

Proof : Using the EM algorithm, [7], to maximize the likelihood of the observable data (X,Y), we iteratively maximize the expectation of the log likelihood of the complete data (X,Y,S)

$$Q(\lambda, \acute{\lambda}) = E[\log(f(X, Y, S|\acute{\lambda})|X, Y, \lambda]$$

The expectation can be replaced by the sum over all possible state sequences S, all possible state mixture densities sequences K, and all possible state transitions mixtures sequences G.

$$Q(\lambda, \acute{\lambda}) = \sum_{S} \sum_{K} \sum_{G} L_\lambda(O, z, S, K, G) \log L_{\acute{\lambda}}(O, z, S, K, G)$$

where

$$L_\lambda(O,Z,S,K,G)=\prod_{t=1}^{T} a_{q_{t-1}q_t}(Z)b_{q_t}(O)$$

Assuming the independence of the observation vectors, and rewriting the previous equation

$$Q(\lambda,\acute{\lambda})=\sum_S\sum_K\sum_G E_{q_tk_tg_tt}\log N(O_t,\acute{\mu}_{q_tk_t},\acute{U}_{q_tk_t})$$

where

$$E_{q_tk_tg_tt}\geq 0$$

This formula is the same as that in [8] and hence the reestimation formulas in (4)-(6) are proven to be the maximum likelihood estimates of the parameters.

Assuming the independence of the observation vectors X and Y, the auxiliary function $Q(\lambda,\acute{\lambda})$ can be written also as

$$Q(\lambda,\acute{\lambda})=\sum_S\sum_K\sum_G H_{q_tk_tg_tt}\log N(Z_t,\acute{\rho}_{q_{t-1}q_tg_t},\acute{V}_{q_{t-1}q_tg_t})$$

where

$$H_{q_tk_tg_tt}\geq 0$$

This formula is the same as that in [8] and hence the reestimation formulas in (1)-(3) are proven to be the maximum likelihood estimates of the parameters.

4. EXPERIMENTS AND RESULTS

The speech is sampled at 16 KHZ, and preemphasized then a Hamming window with a width of 20 ms is applied every 10 ms.

To show the effectiveness of this structure even if the models have the same set of features available for a conventional HMM speech recognition systems, we used a set of features of 12th order LPC-based cepstrum coefficients, and energy and their first difference. No language model is employed and cepstrum mean normalization is used for channel adaptation. 3000 utterances from the TIMIT database are used to train two HMM models: one based on the conventional structure of HMM with feature vectors of length 26, and the new one proposed in this paper. Each phoneme is represented by three states: two of them have feature vector of length 26 while the last one has a feature vector of length 13 (cepstrum coefficients+energy) while the difference vector is associated with the transition probabilities from this phoneme to others. The Gaussian mixtures representing the transition probabilities are tied together into 180 Gaussian probability density functions. The 61 phonemes defined in the TIMIT data base are combined to 48 phonemes for training. The typical 39 distinct phonemes are used as the phoneme set for recognition. An explicit duration model using Gamma probability density function is used within the proposed system at the phonetic boundaries only.

$$f_p(d)=\frac{1}{b_p^{a_p}\gamma(a_p)}d^{a_p-1}e^{-\frac{d}{b_p}}$$

where

$$b_p=\frac{\sum_{i=1}^{J}d_{p,i}^2}{\sum_{i=1}^{J}d_{p,i}}-\sum_{i=1}^{J}d_{p,i},$$

$$a_p=\frac{1}{J}\frac{\left(\sum_{i=1}^{J}d_{p,i}\right)^2}{\sum_{i=1}^{J}d_{p,i}^2-\left(\sum_{i=1}^{J}d_{p,i}\right)^2},$$

J is the number of occurrences of phoneme p in the training data, and $d_{p,i}$ is the duration of the ith occurrence of the phoneme p.

Testing the resulting models using 200 utterances from the TIMIT database, we get a phone recognition accuracy of 58.6% using the conventional HMM and of 62% using the proposed system. Phonemes of short duration like stops which are rarely correctly recognized and most of the time deleted using the conventional system are very rarely missed by the new system. The stop consonant recognition accuracy has increased from 28% to 71% if the phoneme state transitions are trained from manually segmented and labeled data, and to 67% if trained using the extended Baum-Welch training. However, the introduction of the new transition probability models increases the number of substitution errors especially with a similar phoneme. This may be attributed to using only 12 cepstrum coefficients and energy as state-bound observations used in modeling the phonetic units. Increasing the feature vector length for these phonetic units is expected to solve this problem and increase the overall phonetic recognition accuracy.

Figure 1 shows the total recognition accuracy and the stop recognition accuracy for the conventional system, the "inter-state" model tht uses Gaussian mixture models of state transitions at the level of the state, and the "inter-phoneme" model that uses mixture models of state transitions at the level of the phoneme. Mixture Gaussian models were either trained once using manually transcribed phoneme boundaries and not updated during training, or were updated continuously during Baum-Welch training using equations (1)-(3). The significant increase in stop recognition accuracy compared to the total improvement in recognition accuracy

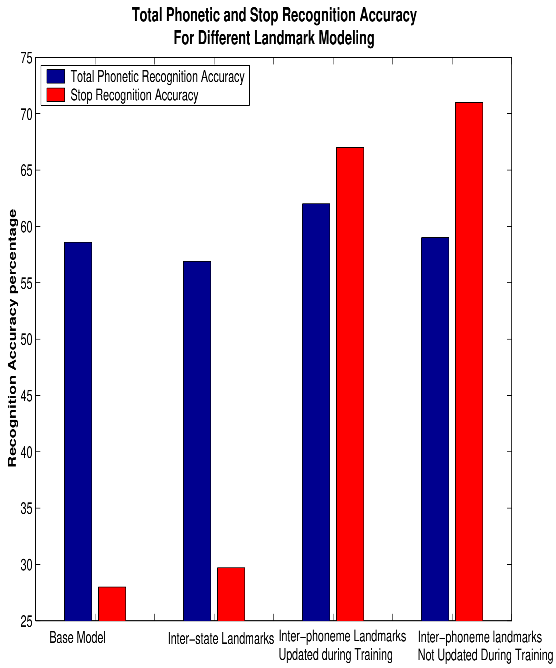

Fig. 1. Total Phonetic and Stop Recognition Accuracy

is due to the ability of the proposed model to precisely determine the start and end of the closure portion of stops and hence to recognize the release portion. This ability is decreased by updating the parameters using the extended Baum-welsh training described before, but the decrease is small compared to the advantage of using training data that is not necessarily labeled and segmented manually.

These results are achieved using the same set of acoustic features and approximately the same number of parameters in both systems. The percentage of correct, substitutions, deletions, and insertions are shown in table 1

Table 1. Phone recognition results for base model and proposed model

	Base Model	**Proposed Model**
Correct	58.6%	62%
Substitutions	30.1%	32.2%
Deletions	11.3%	5.8%
Insertions	13.7%	16.8%

5. DISCUSSION

An improvement of the conventional HMM speech recognition system is introduced. It allows HMM modeling of speech to be more similar to human perception than in the conventional HMM modeling of speech. The advantages of this system include exploiting the information related to temporal correlation in speech, making use of the transition probabilities in the conventional HMM system which were usually negligible during utterance decoding, and allowing the use of heterogeneous acoustic measures for different phoneme transitions. In this work, Cepstrum and energy difference features were used to model the transition probabilities. An extension of this work could be to allow heterogeneous acoustic measures for different state transitions, in order to better exploit the flexibility of the proposed system. This approach is more efficient in consonant recognition and especially stop recognition compared to vowel and nasal recognition.

6. REFERENCES

[1] M. Ostendorf, V. Digalakis, and O. A. Kimball, “Maximum likelihood estimation for multivariate mixture observations of markov chains,” *IEEE Transactions on Information Theory*, vol. 4, no. 5, September 1996.

[2] S. van Vuuren H. Yang and H. Hermansky, “Relevancy of time-frequency features for phonetic classification measured by mutual information,” in *IEEE Proceedings of ICASSP*, 1999.

[3] Stephen Levinson, “Continously variable duration hidden markov models for automatic speech recognition,” *Computer Speech and Language*, vol. 1, 1986.

[4] D. Fohr J. F. Mari and Jean-Claude Junque, “A second-order hmm for high performance word and phoneme-based speech recognition,” *IEEE Transactions on Speech and Audio Processing*, vol. 23, no. 2, March 1996.

[5] Sharlene A. Liu, “Landmark detection for distinctive feature-based speech recognition,” *J. Acoust. Soc. of Am.*, vol. 100, no. 5, November 1996.

[6] S. Levinson B. H. Juang and M. M. Sondhi, “Maximum likelihood estimation for multivariate mixture observations of markov chains,” *IEEE Transactions on Speech and Audio Processing*, vol. IT-32, no. 2, March 1986.

[7] Todd K. Moon, “The expectation maximaization algorithm,” *IEEE Signal Processing magazine*, November 1996.

[8] L. R. Liporace, ” Maximum likelihood estimation for multivariate observations of Markov sources”, *IEEE Transactions on Information Theory*, vol. IT-28, no. 5, pp. 729-734, September 1982.