

A Model For The Acoustic Phonetic Structure Of Arabic Language Using A Single Ergodic Hidden Markov Model

M.A.Mokhtar & A.Z.El-Abdin

Electrical Engineering Department, Faculty of Engineering,
Alexandria University, Alexandria, EGYPT

ABSTRACT

It is proposed to model the acoustic-phonetic structure of the Arabic language using a single ergodic hidden Markov model (HMM), since a single HMM (about 40-50 states) can be used to represent all acoustic phonetic effects. In this paper, we represent the techniques and algorithms used to perform that model, the problems associated with representing the whole acoustic-phonetic structure, the characteristics of the model, and how it performs as a phonetic decoder for recognition of fluent Arabic speech. The model is trained, segmented (manually and automatically), and labeled using a fixed number of phonemes, each of which has a direct correspondence to the states of the model. The model assumes that the observed spectral vectors were generated by a Gaussian source. The inherent variability of each phoneme is modeled as the observable random process of the Markov chain, while the phonotactic model of the unobservable phonetic sequence is represented by the state transition matrix of the HMM. The model incorporated the variable duration feature densities in each state to account for the fact that vowel-like sounds have vastly different duration characteristics than consonant-like sounds. It is shown that the difficulties in developing an acoustic-phonetic model are not due to the inherent deficiencies of the concept presented. Instead they are due to the choice of the phonemes to be modeled, the selected parametrization of the data, and appropriate choice of the variant of the ergodic HMM. The model used for the recognition experiments is clearly not complete, but it adequately performs phonetic transcription of the unknown utterances, thereby serving as the initial step towards continuous speech recognition.

1. INTRODUCTION

Automatic speech recognition has gradually evolved, particularly in the past few years, from limited vocabulary isolated word systems to very large vocabulary, speaker independent recognizers. Most large vocabulary speech recognition systems have adopted the strategy of using a unit of speech that is shorter than a word and for which the number of such units required to represent all spoken sounds is relatively small[1]. Typical choices for such units include diaphones, phonemes, phones, or arbitrary. Each of these choices has advantages and disadvantages in relationship to issues such as the recognition rules' complexity, the size of vocabulary, the speech quality, etc. After studying the phonetic characteristics of modern standard Arabic (few vowels , few consonants), we have chosen phonemes as the basic unit for

representing words[2]. There are several possible choices for what type of signal model is used for characterizing the properties of a given signal. Broadly one can dichotomize the types of signal model into the class of deterministic models, and the class of statistical models[3]. The most popular statistical model used in speech recognition is the Hidden Markov Model (HMM), with the HMM-based system achieving somewhat higher performance than the template based approach[4]. For this procedure, training consists of estimating the parameters (means,covariances) of probabilistic model for each phoneme. A single HMM can be used to represent all acoustic phonetic effects[1] An ergodic HMM was used in which each state represented an acoustic-phonetic unit, and phonotactics are modeled by a simple diagram model. The model incorporated the variable duration feature in each state to account for the fact that vowel-like sounds have vastly different durational characteristics than consonant-like sounds. To classify an unknown utterance, one computes the likelihood that was generated by each of the models derived during training. The utterance is recognized as the phoneme whose gives the highest likelihood score.[5],[1]

2. THE ACOUSTIC-PHONETIC MODEL

A block diagram of a complete continuous speech-recognition system based on HMM is shown in fig.(1). There are essentially three steps in the recognition algorithm [6]:

- 1) Spectral analysis- the speech signal, $s(n)$, is converted to a set of LPC derived cepstral (weighted) and delta cepstral (weighted) vectors.
- 2) Likelihood computation- the sequence of spectral vectors of the unknown speech signal is matched against a set of stored single-phoneme model using the Modified Viterbi algorithm. The output of this process is a set of the likelihood score for all states at all observation time t .
- 3) Decision rule- we can trace back from the final state to recover the highest likelihood score sequence and then an optimal phonetic transcription of the utterance.

2.1. LPC Cepstral Analysis [3],[7],[4],[5]

The speech was recorded using a standard mic, then sampled at 11.1KHz. The speech signal, $s(n)$, is converted to a set of LPC derived cepstral (weighted) and delta- cepstral (weighted) vectors. The LPC from end processing for recognition is shown in Fig.(2). The overall system is a block processing model in which a frame of N samples (30msec) is processed and a vector of features is computed. The steps in the processing are:[3],[4].

1. Pre-emphasis: The digitized speech signal is processed by a first-order digital network in order to spectrally flatten the signal

$$\tilde{s}(n) = s(n) - as(n-1) \quad 0 < a < 1 \quad (1)$$

2. Blocking into frames: Sections of N consecutive speech samples (we use $N=300$ corresponding to 30msec) of signal are used as a single frame. Consecutive frames are spaced M samples apart (we use $M=100$ corresponding to 10 msec , or 20 msec overlap)[3].
3. Frame windowing: Each frame is multiplied by an N -sample window (we use Hamming window) so as to minimize the adverse effects of chopping an N -sample section out of the speech signal.
4. Auto correlation analysis: Each windowed set of speech samples is auto-correlated to give a set of $(p+1)$ coefficients, where p is the order of the desired LPC analysis (we use $p=8$)[7].
5. LPC /Cepstral analysis: For each frame, vectors of LPC coefficients are computed from the auto-correlation vector using Levinson or Durbin recursion method. The LPC derived cepstral vector is then computed up to the Q 'th component, where $Q > p$, $Q=12$ in our implementation[3],[7]
6. Cepstral weighting: The Q -coefficient cepstral vector, $C_l(m)$, at time frame l , is weighted by window $W_c(m)$ of the form

$$W_c(m) = \left[1 + \frac{Q}{2} \sin\left(\frac{\pi m}{Q}\right) \right] \quad 1 \leq m \leq Q \quad (2)$$

$$\text{to give } \tilde{C}_l(m) = C_l(m) \cdot W_c(m) \quad (3)$$

7. Delta cepstrum: The time derivative of the sequence of weighted cepstral vector is approximated by a first-order orthogonal polynomial over a finite length window of $(2K+1)$ frames, centered around the current vector ($K=2$ in our implementation); hence, the derivative is computed from a 5-frame window. The derivative contains important information about the delta temporal rate of change of the cepstrum, and is computed as:

$$\Delta C_l(m) = \left[\sum_{k=-K}^K k C_{l-k}(m) \right] \cdot G \approx \frac{\delta C_m}{\delta t}, 1 \leq m \leq Q \quad (4)$$

Where G is a given term so that the variances of

$C_l(m)$ and $\Delta C_l(m)$ are about the same (For our system, the value of G was 0.375)[5].

The overall observation vector, O_l , used for scoring the HMM's is the concatenation of the weighted cepstral vector, and the corresponding weighted delta cepstrum vector, i.e.:

$$O_l = \{C_l(m), \Delta C_l(m)\} \quad (5)$$

and consists of 24 coefficients per vector, which were used in all the experiments described below.

2.2. HMM Characterization Of Phonemes [1]

The model that we use to represents the acoustic-phonetic structure of the Arabic language is the Continuously Variable Duration HMM (CVDHMM). We have considered the special

case of ergodic or fully connected HMMs in which every state of the model could be reached (in a single step) from every other state of the model. This type of model has the property that every

a_{ij} coefficient is positive[1]. The states of the model $\{q_i\}_{i=1}^n$ represent the hidden phonetic units. The phonotactic structure of the language is modeled, to a first order approximation, by the state transition matrix a_{ij} , which defines the probability of occurrence of state (phoneme) q_i at time $t+\tau$ conditioned on state (phoneme) q_j at time t , where τ is the duration of phoneme i . The information about the temporal structure of the hidden units is contained in the set of durational densities $\{d_j(t)\}$. The speech acoustic correlates are the observations, denoted O_t , and their distributions, which are defined by a set of observation densities $\{b_j(O_t)\}_{j=1}^n$. Such a model is illustrated in Fig (3), in which

each state j is characterized by the following:

1. A State transition vector: a_j , with components a_{ij} = probability of making a transition to state i (at the next transition instant), given that the system is currently at state j . For the ergodic model of Fig, all states can be transmitted to all other states, and satisfies the stochastic constraint:

$$\sum_{i=1}^n a_{ji} = 1, a_{ji} > 0, \forall 1 \leq j, i \leq n \quad (6)$$

2. A State observation density, $b_j(O_t)$, of the form [4],[5],[8]:

$$b_j(O_t) = \sum_{m=1}^M C_{mj} N[O_t, \mu_{jm}, U_{jm}] \quad (7)$$

i.e., a continuous mixture density where O_t is the observation vector (e.g., cepstral coefficient vector resulting from the LPC analysis), C_{jm} is the mixture weight for the m 'th component in state j , $N[\cdot]$ represents a multivariate normal density [i.e., Gaussian] with the mean vector μ_{jm} for mixture m in the state j , and covariance matrix U_{jm} for mixture m in state j . Typically, we use anywhere from $M=1$ to $M=q$ mixture components. The mixture gains C_{jm} satisfy the stochastic constraint:

$$\sum_{m=1}^M C_{jm} = 1, C_{jm} > 0, 1 \leq j \leq n, 1 \leq m \leq M \quad (8)$$

So that the probability density function (pdf) is properly normalized, i.e.,

$$\int_{-\infty}^{\infty} b_j(x) dx = 1, \quad 1 \leq j \leq n \quad (9)$$

In practice, we have observed that components of O are essentially uncorrelated. Hence, we assume that all components of O are statistically uncorrelated. Thus, U_{jm} becomes a diagonal covariance matrix, and (6) can be expressed simply as [4]:

$$b_j(O_t) = \sum_{m=1}^M \frac{\prod_{d=1}^D \exp[-(O_t(d) - \mu_{jmd})^2 / 2\sigma_{jmd}^2]}{(2\pi)^{D/2} (\prod_{d=1}^D \sigma_{jmd}^2)^{1/2}} \quad (10)$$

Where $O_t(d)$ is the d 'th component of the observation vector, D is the number of components in O_t , μ_{jmd} is the

d 'th component of the μ_{jm} , and σ_{jmd}^2 is the d 'th covariance of U_{jm} .

3. State duration probability [5], $d_j(\tau)$, where τ is the number of frames spent in state j , and d_j is a parametric state duration density. In practice, it has been argued that the duration structure of the speech code is best modeled by a set of asymmetrical Gamma densities with:

$$d_j(\tau) = \frac{\zeta_j^{\nu_j} d^{\nu_j-1} e^{-\zeta_j \tau}}{\Gamma(\nu_j)} \quad (11)$$

with parameters $\nu_j = \text{mean}^2/\text{var}$. and $\zeta_j = \text{mean}/\text{var}$.

Based on the above, the process of building an HMM of the type shown in Fig, to characterize a phoneme, requires estimation of:

- 1) N^2 values of a_{ij} , the state transition coefficients;
- 2) NM values of C_{mj} , the mixture gains;
- 3) NMD values of $\mu_{mj d}$, the mean values of the observations;
- 4) NMD values of $[U_{jm}]_d$, the diagonal covariances of the observations;
- 5) N values of mean ($d_j(\tau)$), the state-duration density mean;
- 6) N values of variance ($d_j(\tau)$), the state-duration density variance.

All these parameters are estimated or measured directly from a training set as discussed in the next section.

2.3. Parameter Estimation

The model parameters are estimated from the analyzed speech and statistics of the phonetic segments (manually) except for the state transition matrix A , which is estimated from a very large corpus of text (novel). The count of the phoneme transitions was obtained directly from the transcriptions used in that novel. Each of the words in the novel was considered an isolated utterance, bounded by silence, and transitions between the phonemes were counted. The obvious shortcoming of such a procedure is that the frequency of occurrence of each of the words was not included in the estimation[1]. A study of the relationships between the different sounds of modern standard Arabic enabled us to arrive at a model which consists of some phonetic elements which are most suited to the recognition of speech[2]. There are six Arabic vowels, three short and three long vowels. In Arabic, vowels cannot be initials and can occur either between two consonants or final in word. Unlike the consonants, the vowels have no segment duration charges, except that a long vowel is about twice the length of a short vowel[2]. There are 28 consonants in Arabic and all can occur as initial, intervocalic, or syllable closing. Intervocalic and initial consonants have durations which are about half those of syllable closing or syllable suffix consonants. Some continuant initial consonants can be large than intervocalic consonants, but the difference in length is not very significant[2]. The resulting set of 40 phonemes and silence which formed the basis of the 41 states (phonemes) of the acoustic-phonetic model are listed in [2]. The CVDHMM requires information about the distributions of duration's of each of the phonemes (states). The training data were segmented manually into phonemes, then the duration of each phoneme was measured from the direct segmentation. Then, the duration parameters are obtained by calculating the mean duration of each phoneme and its variance,

which are then converted into the parameters μ and ζ . However, for the B parameters, experience has shown that good initial estimates are essential in the continuous distributions case. Such initial estimates can be obtained in a number of ways, including manual segmentation of the observation sequences into phonemes (states) with averaging of observations within states[5]. All frames (vectors) for a given state (phoneme) are used as input to a clustering algorithm (i.e., a vector quantizer design procedure[9][10]) which determines the best M cluster solution (using an Euclidean distortion measure), which is generally the centroid of the frames in the training set assigned to the m 'th region. From the clustering, an updated set of model parameters are derived as follows[3]:

C_{jm} = number of vectors classified in cluster m of state j divided by the number of vectors in state j .

μ_{jm} = sample mean of the vectors classified in cluster m of state j ,

$$\mu_{jm}(d) = \frac{1}{T} \sum_{t=1}^T O_t(d) \quad (12)$$

U_{jm} = sample covariance matrix of the vectors classified in cluster m of state j ,

$$U_{jm}(r, s) = \frac{1}{T} \sum_{t=1}^T (O_t^{(r)} - \mu_{jm}(r))(O_t^{(s)} - \mu_{jm}(s)) \quad (13)$$

Because there was only a limited amount of data available, a constant ($\cong 0.02$ or 0.07) was added to the diagonal elements of the covariance matrix, U_{jm} , to prevent it from becoming singular. This method had a great effect on recognition results[1].

3. THE RECOGNITION EXPERIMENT

In the case of a single HMM all we need to determine is the best estimate of the state sequence of the model which coincidentally provides the best estimate of the phoneme string of the utterance. An efficient way to optimally determine a state sequence of a CVDHMM through an utterance is using a modified Viterbi algorithm which needs to account for the durational densities. This calculation is implemented as follows[1][5]. Let $d_{\tau}(j)$ be the likelihood of the state sequence ending in state j and corresponding to the first τ observations which maximize the joint likelihood of state and observation sequence :

$$\alpha_{\tau}(j) = \max_{1 \leq i \leq n} \{ \max_{\tau \leq t} \{ \alpha_{t-\tau}(i) a_{ij} d_j(\tau) \prod_{\theta=1}^{\tau} b_j(O_{t-\tau+\theta}) \} \}$$

for $1 < j < n$ and $1 < t < T$. If, at the same time, we set

$$\beta_{\tau}(j) = (i, \tau) = \arg \max_{i, \tau} \{ d_{\tau}(j) \}$$

then we can trace back from the final state to recover the optimal state sequence and then an optimal phoneme sequence. Thus we obtain the phonetic transcription of the utterance.

$$j_{\tau} = \arg \max_j \{ \alpha_{\tau}(j) \}$$

4. RESULTS AND CONCLUSION

Two test sentences were used for testing the model. The sentences contain mostly the phonemes of the Arabic language

