

SPEECH RECOGNITION USING SEMI-HIDDEN MARKOV MODELS OF MULTIPLE FEATURES

X. Zhang* and J.S.D. Mason*

ABSTRACT

Semi-hidden Markov models (SHMMs) have been suggested and applied to isolated speaker-dependent E-set recognition. The SHMM differs from the conventional hidden Markov model (HMM) in that its states can be classified into types. A function which detects signals corresponding to state types is thus included in the SHMMs and utilized to supervise the estimation of their parameters. This general structure is implemented in the recognition experiment as models with their states classified into stationary and transient types. The average recognition error rate is about 18.9% which compares favourably with the average of about 36.4% reported when using a dynamic time warping (DTW) recognition system by Lienard and Soong (ref 3) on an equivalent vocabulary. Tests using corresponding HMMs show similar results to that of the DTW system.

INTRODUCTION

To date, almost all known applications of the Markov speech modeling variations assume that the underlying Markov chain is hidden, i.e., it is not directly observable (ref 1). As a result, it is the model parameter estimation procedure that segments the input signals into quasi-stationary intervals, associates them with different states, and identifies their corresponding parameter distributions when the model is trained. In this sense, the HMM is uncontrollable. Such an approach works reasonably well for speech recognition such as reported in (ref 2). However, like most pattern recognition approaches to isolated word recognition, it runs into difficulties for confusable vocabulary like the English E-set {b,c,d,e,g,p,t,v}.

A new proposal is made by Zhang and Mason (ref 6) to take a degree of the above mentioned control over and to explicitly associate certain types of signal with specific state types and to some extent freely distribute model states across different linguistic units. It is seen beneficial in that, by this means, some extra human knowledge about the speech signal can then be adopted to supervise the automatic process of Markov modeling.

To distinguish the E-set words, it is important to emphasize the differences of their initial transients, and the formant evolutions at the beginnings of some vowel segments can also be important features (ref 3). Explicitly modeling the transients seems to be a suitable Markov approach to emphasizing the importance of these transients. The conventional HMMs are seen to be unsuitable for this task because a transient is usually too short to be easily caught. By introducing supervision in model training, SHMMs could arrange several states as wished for a transient and give the corresponding broad spectral evolution if a reliable feature can be found.

SHMM

Towards introducing some sort of supervision into the Baum-Welch Markov model training algorithm (ref 4), the SHMM suggested here assumes that every state has its own type and only signals with certain known characteristics coinciding with the state type can be generated there. Correspondingly, the SHMM includes a detector which extracts the predefined characteristics of the training inputs to

*Dept. of Elec. Eng., University College of Swansea, Swansea SA2 8PP

supervise the estimator. Coarsely distinguishing some simple characteristics (signal type) such as voiced vs. unvoiced and transient vs. quasi-stationary does not present too much difficulty. The detection does not have to be very accurate because the estimator can somehow judge by using other knowledge sources (e.g., the grammar constraint conveyed in the model state transition matrix of the left-to-right model). The idea is implemented essentially by extending the original transition probability, a_{ij} , to $a_{ij} \cdot e(q_i, q_j, s)$ where the $e()$ refers to the positive valued detection function and q_k stand for the type of state k . The $e()$ hence acts as a weighting function of the conventional transition probability a_{ij} according to whether or not the observed signal s fits into the category required by the state types. The explicit form of the detection, however, is application dependent.

In addition, it is also assumed to be possible for an SHMM to accommodate separately more than one parallel independent observation source. The overall observation density is the product of the Gaussian mixture densities of each individual observation source.

The training of SHMM is done by modified Baum-Welch algorithm (ref 7). The standard HMM may be regarded as a special case of the SHMM, if a degenerated detection function $e() \equiv 1$ is assumed.

E-SET RECOGNITION

An attempt has been made to apply the SHMMs to speaker-dependent isolated word recognition of the English E-set. To explicitly modeling the important E-set transients, SHMM states are classified as transient and quasi-stationary types. Two illustrative SHMM models for utterances "b" and "c" are given in Fig 1. Using two feature sequences m_t and n_t (described later), the $e()$ function is made to reflect some expected transition properties between the two signal types. For example, the $e()$ returns a relatively large value when entering a transient state only if the local interframe difference is big and again a large value when entering a stationary state only if the local interframe difference is small.

To extract feature from speech signals, two degrees of overlap of the 20ms Hamming windowed frames are used: one is a large degree of overlap of 83.3% so that analyses can be made often enough to catch the spectral contour of transients and, also, obtain more spectral samples for statistical estimation. The other is 50% used for quasi-stationary sounds. The two overlap rates result in a bias to the transients similar to the strategy suggested in (ref 3) for emphasizing transient signals. The second overlap rate of 83.3% was chosen (somewhat arbitrarily) not as an optimum but merely as a high rate to demonstrate the principle. The switching between the two overlap rates is achieved by detecting the regression sequence m_t (defined later) to be above or below a preset threshold.

Cepstral coefficients derived from linear predictive analysis (LPC-cepstra), c_i , are used as one feature source. Linear regression analysis is applied to the $k \cdot c_t^{(k)}$ sequence for each fixed k to get a second feature sequence $y_t^{(k)}$ which is thought to be largely independent of the c_i ($c_t^{(k)}$ denotes the c_k at time t). It is not difficult to see that the $y_t^{(k)}$ are averaged differences (or derivatives equivalently) of the input sequence $k \cdot c_t^{(k)}$. They thus approximate the local differences of the root-power sums (ref 5) and indicate spectral differences (also formant evolution). Each $y_t^{(k)}$ is extracted from a frame of 70ms of speech which is thought to be able to reflect the local broad spectral evolution for most cases. For a relatively short transient interval, this feature should be quasi-stationary. This is why it is adopted as the second feature source. The summations of the squared $y_t^{(k)}$, giving the feature m_t , and short term linear regressions of the m_t , n_t , are used as the input of the detection function $e()$.

It is known that the expectation of duration of, say, state i can be derived easily from

$$d_i = 1 / (1 - a_{ii}). \quad (1)$$

Similar to the approach presented in (ref 2), additional simple post-processors of Poisson duration models, whose duration expectations are equal to d_i derived from Equation (1), are included.

The pattern matching is generally scored by

$$\log P = \log P + \frac{\alpha}{N} \sum_{j=1}^N \log [p(l_j/T)] + \frac{\beta}{N_t} \sum_{j=1}^{N_t} \log [p(l'_j/T)] \quad (2)$$

where the P is the probability density of matching derived from the conventional Viterbi process, the $p(x)$ is the Poisson density function with duration expectation x and the l_j/T and l'_j/T are the normalized times spent in j th encountered state or state type respectively, along the optimal Viterbi alignment path.

Five model structures are investigated. They are

M1: HMM, trained by LPC-cepstrum feature source only, $\alpha = 100$, $\beta = 0$;

M2: HMM, trained by both feature sources, $\alpha = 100$, $\beta = 0$;

M3: SHMM, trained by both feature sources, $\alpha = 100$, $\beta = 0$;

M4: SHMM, trained by both feature sources, $\alpha = 0$, $\beta = 100$;

M5: SHMM, trained by both feature sources, $\alpha, \beta = 100$.

Extensive Recognition experiments on E-set tokens of 9 speakers (5 male, 4 female) using these systems have been performed. Average recognition performance is reported in Fig 2; the utterance matching is done in two ways: matching only the LPC-cepstrum vector sequences (dashed line) and matching both feature sources (solid line).

The average error rate is only about 15.3% using M5 and two-feature-matching. The M2 achieves better performance than M1 when using two-feature-matching process. But the one-feature-matching process of M2 generally gives the same results as that of M1. When SHMMs are used, recognition error rates decrease sharply even for the one-feature-matching. The average recognition error rates are 31.8%:22.6% and 24.7%:18.4% (M2:M3) using one-feature-matching and two-feature-matching approaches respectively.

On average, introducing the state-type-based log-duration model gives slightly better recognition performance than the state-based (for example, the average recognition error rate of M4 is about 1.0% lower than that of M3). It means that state-type-based duration is more consistent, further justifying the idea of state type classification of the speech signals and the application of SHMM.

An almost equivalent vocabulary {p,b,t,d,v,z} is used by Lienard and Soong (ref 3) with an LPC template dynamic time warping (DTW) recognition system. Results are compared with our (S)HMMs systems using the vocabulary {p,b,t,d,v,e} in Fig 3. It is clearly seen that the error rates of the LPC/DTW system are equivalent to that of conventional HMMs of M1, but SHMMs of M5 do much better (18.9% (M5) vs. 36.4% (LPC/DTW)).

More relevant details are seen in (ref 7).

CONCLUSION

A Semi-hidden Markov model has been described and its performance on speaker-dependent isolated English E-set recognition is reported. It is shown that the SHMM can give much better E-set recognition performance than conventional hidden Markov model (HMM) and better than any comparable results reported. It

may be concluded that the proposed semi-hidden Markov model is useful to speech processing in that it provides a means of incorporating high-level human knowledge (in the form of signal type) into the automatic process of modeling target speech signals, making the underlying Markov process less-hidden.

REFERENCES

1. S.E. Levinson et al., "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Processes to Automatic Speech Recognition," Bell Syst. Tech. J., Vol. 62, No. 4, 1035 (1983).
2. L.R. Rabiner et al., "Recognition of Isolated Digits Using Hidden Markov Models with Continuous Mixture Densities," At&T Tech. J., Vol. 64, No. 6, 1211 (1985).
3. J. Lienard and F.K. Soong, "On the Use of Transient information in Speech Recognition," Proc. ICASSP '84, 17.3.
4. B.H. Juang, "Maximum-Likelihood Estimation for Mixture Multivariate Stochastic Observations of Markov Chains," AT&T Tech. J., Vol. 64, No. 6, 1235 (1985).
5. B.A. Hanson and H. Wakita, "Spectral Slope Based Distortion Measures for All-Pole Models of Speech," Proc. ICASSP '86, 757.
6. X. Zhang and J.S. Mason, "Modified Markov Model for Speech Recognition," Proc. 18th JAACE Symposium on Stochastic Systems Theory and Its Applications, Tokyo, 257 (1986).
7. X. Zhang, "A Semi-Hidden Markov Model and Its Application to Speech Recognition," Ph.D. Thesis, submitted to University College of Swansea, U.K.

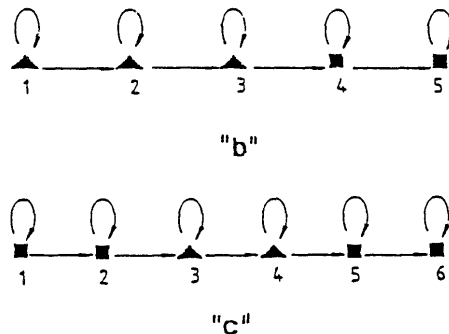


Fig 1 SHMM Models for "b" and "c" (■ :stationary state; ▲ :transient state)

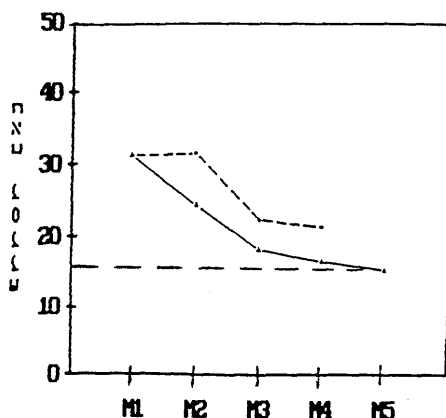


Fig 2 E-set recognition performance

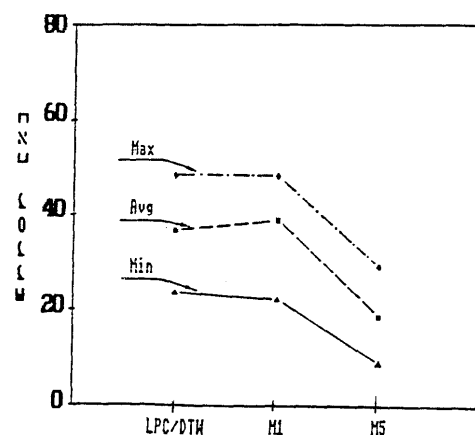


Fig 3 Error comparison of 3 systems (LPC/DTW (ref 3), M1 and M5)