A FACTORIAL HMM APPROACH TO SIMULTANEOUS RECOGNITION OF ISOLATED DIGITS SPOKEN BY MULTIPLE TALKERS ON ONE AUDIO CHANNEL

Ameya Nitin Deoras and Mark Hasegawa-Johnson

University of Illinois at Urbana-Champaign
deoras@uiuc.edu, jhasegaw@uiuc.edu

ABSTRACT

This paper addresses the novel problem of recognizing digits spoken simultaneously by two different talkers. A Factorial Hidden Markov Model architecture is proposed to accurately model the simultaneous utterance of two digits. Nadas' MIXMAX approximation is extended to a mixture of Gaussians observation PDF which enables the implementation of the proposed system. The multiple digit recognizer is found to successfully recognize pairs of simultaneous utterances of digits at 0db SNR with up to 89% accuracy.

1. INTRODUCTION

The idea of the Factorial Hidden Markov Model (FHMM) was first developed by Ghahramani as an alternative to traditional HMMs [1]. It has been shown that factorial HMMs are better suited to model loosely coupled random processes [1], [2]. Furthermore, efficient algorithms for the estimation of parameters of FHMMs have also been developed [1]. Our approach is, however, a little different. We use the FHMM architecture to combine two existing HMMs of two independent random processes, i.e. the simultaneous utterance of two digits. Since the isolated digit HMMs have already been trained, no additional training of the FHMM is required. Roweis showed that an FHMM can be used in such a way to model audio signals from different sources for a computational auditory scene analysis application [3]. We build on Roweis' method for the recognition of two digits spoken simultaneously by two speakers.

The motivation behind this model arose from the observed interaction of the log spectra of two signals that are additive in the time domain. Nadas et al. have shown that an additive combination of two sound signals (or a signal and noise) $\bar{Y}(j\omega) = \bar{X}(j\omega) + \bar{Z}(j\omega)$ can be accurately modeled by the element-wise maximum of their log magnitude spectra. This is referred to as the MIXMAX approximation [4].

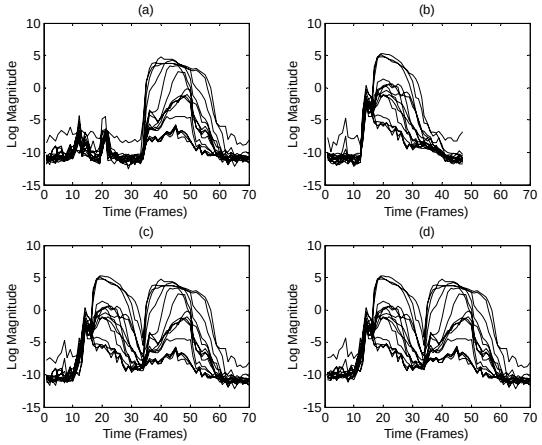

Figure 1. Log spectrum MIXMAX approximation; (a) MFSC observation sequence of the digit 'ONE'; (b) MFSC observation sequence of the digit 'TWO'; (c) the element-wise maximum of sequences (a) and (b); (d) the MFSC observation sequence generated from the addition of the utterances of the digits 'ONE' and 'TWO' in the time domain.

$$\log|\bar{Y}(j\omega)| \approx \max(\log|\bar{X}(j\omega)|, \log|\bar{Z}(j\omega)|) \quad (1.1)$$

The MIXMAX approximation also holds for Mel Frequency Spectral Coefficients (MFSC) [5] as shown in Figure 1.

In the following section, we present a short discussion on the architecture of the factorial HMM, its topological equivalence to an HMM and how its parameters can be estimated given the parameters of the two HMMs it is composed of. We also present our extension of the MIXMAX output probability density result to a mixture of Gaussians PDF. In sections 3 and 4, we describe the implementation and testing of the simultaneous multiple digit recognition system followed by a discussion of its performance in section 5.

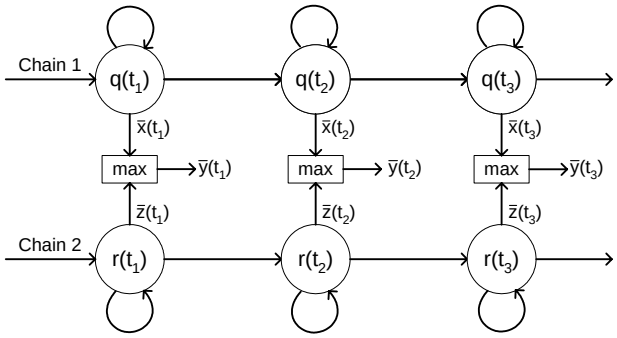

Figure 2. The combination of two HMM chains into an FHMM structure.

2. THE FACTORIAL HIDDEN MARKOV MODEL

We use a factorial HMM architecture with the same structure as that proposed by Roweis [3]. The FHMM can be visualized as a pair of hidden-state Markov chains that evolve independently of each other. Each of these states produces an independent observation vector at each time step. The output of the FHMM in every frame is simply the maximum of the two outputs proposed by each chain as described by Equation (1.1) and illustrated in Figure 2.

2.1. Topological Equivalence to an HMM

Consider a factorial HMM with two chains (denoted by a superscript index) containing Q and R states respectively. This FHMM can be shown to be topologically equivalent to an HMM with $Q \times R$ states [2]. The transition matrix for such an HMM is given by

$$a^{FHMM}(i, j \to k, l) = a^1_{i \to k} \times a^2_{j \to l} \quad \begin{matrix} 1 \le i, k \le Q \\ 1 \le j, l \le R \end{matrix} \qquad (2.1.1)$$

where the states of the FHMM are indexed by the pair of state indices of chains 1 and 2.

2.2. The Output Probability Distribution

Methods for calculating the posterior probability of factorial HMMs have been formulated by Logan and Ghahramani to overcome the large computations involved [1], [2]. Roweis uses a log probability upper bound for a similar reason to compute the best joint state trajectory. We propose a simpler approach that takes advantage of the MIXMAX approximation as well as the efficient recognition algorithms that are already available for HMMs. To make this possible, we derive the output probability distribution of the FHMM's equivalent HMM. This is done by extending the results proposed by Nadas et al. to the case of a mixture of Gaussians observation PDF.

Let the state indices of the two independent HMM chains (denoted by a superscript index) that compose the FHMM be $q(t)$ and $r(t)$, and let the proposed MFSC observation vectors be $\bar{x}(t)$ and $\bar{z}(t)$ respectively. The output of the FHMM is given by,

$$\bar{y}(t) = \max(\bar{x}(t), \bar{z}(t)) \qquad (2.2.1)$$

where $\max(\bar{x}(t), \bar{z}(t))$ is the element-wise maximum.

Since the two processes are independent, it follows from Equation (2.2.1) that,

$$F_{\bar{y}}(\bar{\lambda}) = F_{\bar{x}}(\bar{\lambda}) F_{\bar{z}}(\bar{\lambda}) \qquad (2.2.2)$$

where $F_{\bar{y}}(\bar{\lambda}) = P(\bar{y} < \bar{\lambda}) = \int_{-\infty}^{\bar{\lambda}} p_{\bar{y}}(\bar{y}) d\bar{y}$ is the CDF of $\bar{y}$.

Differentiating Equation (2.2.2) gives us the PDF of $\bar{y}(t)$.

$$p_{\bar{y}}(\bar{\lambda}) = p_{\bar{x}}(\bar{\lambda}) F_{\bar{z}}(\bar{\lambda}) + p_{\bar{z}}(\bar{\lambda}) F_{\bar{x}}(\bar{\lambda}) \qquad (2.2.3)$$

which is the same result as that proposed by Nadas et al. [4].

2.3. Extension to a Mixture of Gaussians

Since each HMM state has an output probability density function represented by a mixture of Gaussians, we can write,

$$\begin{aligned} b^1_q(\bar{x}_t) &= \sum_{m=1}^{M} c^1_{q,m} N(\bar{x}_t \mid \bar{\mu}^1_{q,m}, \Sigma^1_{q,m}) \\ b^2_r(\bar{z}_t) &= \sum_{m=1}^{M} c^2_{r,m} N(\bar{z}_t \mid \bar{\mu}^2_{r,m}, \Sigma^2_{r,m}) \end{aligned} \qquad (2.3.1)$$

$$N(\bar{o} \mid \bar{\mu}_m, \Sigma_m) = \frac{1}{\sqrt{(2\pi)^n \mid \Sigma_m \mid}} e^{-\frac{1}{2}(\bar{o}-\bar{\mu}_m)' \Sigma_m^{-1} (\bar{o}-\bar{\mu}_m)}$$

where M is the number of Gaussians in each mixture, c is the mixture coefficient and n is the dimensionality of the output vectors.

Extending the results of Equation (2.2.3) to the HMM output PDFs defined in Equation (2.3.1), the FHMM output probability density function can be written as,

$$b_{q,r}(\bar{y}_t) = b^1_q(\bar{y}_t) \int_{-\infty}^{\bar{y}_t} b^2_r(\bar{z}_t) d\bar{z}_t + b^2_r(\bar{y}_t) \int_{-\infty}^{\bar{y}_t} b^1_q(\bar{x}_t) d\bar{x}_t \qquad (2.3.2)$$

where the superscripts denote the chain index.

The integrals in Equation (2.3.2) are the cumulative density functions of $b_q^1(\overline{x}_t)$ and $b_r^2(\overline{z}_t)$. The d-variate Gaussians are assumed to be diagonal covariance Gaussians to reduce the order of computation from O(n^d) to O(nd). This assumption enables us to represent the multivariate Gaussian as the product of d univariate Gaussians. Therefore, Equation (2.3.1) can be written as,

$$b_q^1(\overline{x}_t) = \sum_{m=1}^{M} c_{qm}^1 \prod_{p=1}^{n} \frac{1}{\sigma_{qm,p}\sqrt{(2\pi)}} e^{-\frac{1}{2}\left(\frac{x_{t,p}-\mu_{qm,p}}{\sigma_{qm,p}}\right)^2} \quad (2.3.3)$$

where $\sigma_{qm,i}^2$ is the element at position *(i,i)* on the diagonal of the covariance matrix Σ_{qm}.

The integral of the diagonal covariance Gaussian mixture is therefore given by,

$$\int_{-\infty}^{\overline{y}_t} b_q^1(\overline{x}_t)d\overline{x}_t = \int_{-\infty}^{\overline{y}_t} \left(\sum_{m=1}^{M} c_{qm}^1 \prod_{p=1}^{n} \frac{1}{\sigma_{qm,p}\sqrt{(2\pi)}} e^{-\frac{1}{2}\left(\frac{x_{t,p}-\mu_{qm,p}}{\sigma_{qm,p}}\right)^2} \right) d\overline{x}_t \quad (2.3.4)$$

The integral of the sum of the Gaussians is equivalent to the sum of the integral of the Gaussians. Also, since we are representing an *n*-variate Gaussian as a product of *n* univariate Gaussians, the integral to $\overline{y}_t$ of the product of Gaussians should be equal to the product of the integrals of each Gaussian to $y_{t,p}$, where $\overline{y}_t =< y_{t,1}, y_{t,2}, ..., y_{t,n} >$. Therefore, we get the following result for the CDF of *n*-variate Gaussians,

$$\int_{-\infty}^{\overline{y}_t} b_q^1(\overline{x}_t)d\overline{x}_t = \sum_{m=1}^{M} c_{qm}^1 \prod_{p=1}^{n} \left(\int_{-\infty}^{y_{t,p}} \frac{1}{\sigma_{qm,p}\sqrt{(2\pi)}} e^{-\frac{1}{2}\left(\frac{x_{t,p}-\mu_{qm,p}}{\sigma_{qm,p}}\right)^2} dx_{t,p} \right) \quad (2.3.5)$$

In other words, the CDF of every *n*-variate diagonal covariance Gaussian is just the product of the CDFs of the *n* univariate Gaussians that it is composed of.

Equations (2.3.2), (2.3.3) and (2.3.5) alone completely specify the output probability density of the FHMM.

3. RECOGNITION SYSTEM

The baseline isolated digit speech recognition system was built in Matlab with the help of routines provided by Kevin Murphy in the Hidden Markov Model Toolbox for Matlab [8]. Each model was designed with 8 states and a mixture of 120 Gaussians per state [6]. To generate the observation vectors, a 32ms frame of data was used to compute 20 Mel Frequency Spectral Coefficients. Adjacent frames of MFSC vectors were concatenated to produce the observation sequence.

Each digit model was trained on 100 utterances by 50 male speakers from the NIST/TIDIGITS speech corpus. On clean speech, the baseline system performed with a word recognition error rate of 3%.

4. EXPERIMENTS

In this section, we evaluated the performance of the simultaneous multiple-digit recognition system on two distinct recognition tasks described in sections 4.1 and 4.2. In each test, two utterances (produced by different speakers) from the TIDIGITS corpus were combined at 0db SNR in the time domain, followed by the computation of an MFSC observation sequence. The recognition system was then presented with the task of finding the Factorial HMM (out of all combinations of allowed digits) that best modeled the mixed utterance. The performance of the system was then compared with the performance of the baseline system on the same task.

4.1. Recognition of 'Signal' Digit Given 'Interference' Digit

In the first set of experiments, one of the digits of the double-digit pair was treated as the 'signal' and the other as a known 'interference' at 0db SNR. The system's performance in recognizing the signal digit, given the knowledge of the interference digit, was tested for different combinations of utterances of digits.

From the results, presented in Table 4.1, we can see that the multiple-digit recognition system showed an average relative improvement in word accuracy of 35% over baseline.

4.2. Simultaneous Double-Digit Recognition

In the second set of experiments, the system was not given any prior information on either digit and its simultaneous double-digit recognition performance was evaluated. A successful recognition of both digits was considered a 'complete success' (CS) and a recognition of one of the digits in the pair was considered a 'partial success, partial failure' (PSPF). The average word accuracy per channel (recognition rate) was computed by,

$$\text{Recognition Rate (\%)} = \frac{CS + 0.5 \times PSPF}{N} \quad (4.1)$$

where N is the total number of recognitions performed.

Table 4.2 lists the results from the double-digit recognition tests. The multiple digit recognition system showed an average relative improvement in word accuracy of 105% over baseline.

TABLE 4.1

Average recognition rates of the 'signal' digit, given the 'interference' digit. S is the set of allowed digits; ID is the interference digit; N is the total number of recognitions or trials; BRR is the baseline system recognition rate; P represents the case where FHMMs of pairs of the same digit are allowed (i.e. 'signal' and 'interference' can be the same digit); NP represents the case where such models of pairs of the same digit are not allowed.

S	ID	N	BRR	FHMM Average Word Accuracy	
				P	NP
0 - 5	4	6	67%	83%	100%
0 - 8	3	45	55%	64%	68%

TABLE 4.2

Average recognition rates for the task of simultaneous double-digit recognition. N is the total number of recognitions or trials assuming no pairs of the same digit. The CS/N column represents the 'complete success' recognition rate assuming no models of pairs of the same digit.

S	N	BRR	$\frac{CS}{N}$	FHMM Average Word Accuracy per Channel	
				P	NP
5 - 8	6	67%	100%	-	100%
4 - 8	40	36%	78%	-	89%
1 - 5	50	38%	70%	78%	84%

5. DISCUSSION

It is interesting to note that while the performance of the baseline system dropped in the second task, the performance of the double digit recognizer increased. This may be due to the fact that since the FHMM is designed to model any combination of digits, it produces a higher overall accuracy in modeling a more diverse set of digit combinations.

Another observation that can be made from the results is that the system cannot easily model simultaneous utterances of the same digit by different speakers. This is because in the FHMM structure, each chain is competing to explain the observation of the same digit, resulting in a drop in performance.

The multiple digit recognition system, besides being able to recognize simultaneous utterances, can also be used as a technique for robust speech recognition in non-stationary noise. It has a similar range of word accuracy at 0db SNR when compared to current highly robust speech recognizers [7]. However, while most standard methods for speech recognition in noise (e.g. spectral subtraction, Wiener filtering) assume stationary or slowly-varying background noise, the FHMM approach is robust for noise that is rapidly varying over a large dynamic range, like speech or music.

6. CONCLUSIONS

A factorial HMM modeling approach for the simultaneous recognition of multiple digits has been presented. The MIXMAX algorithm was extended to include mixtures of Gaussians which enabled the implementation of a simultaneous multiple digit recognition system. The system was shown to have significant success in double-digit recognition at 0db SNR.

7. REFERENCES

[1] Z. Ghahramani and M.I. Jordan, "Factorial Hidden Markov Models," *Machine Learning*, 29, pp. 245-275, 1997.

[2] B. Logan and P. Moreno, "Factorial HMMs for Acoustic Modeling," *ICASSP*, pp. 813-816, 1998.

[3] S.T. Roweis, "One Microphone Source Separation," *Neural Information Processing Systems* 13, pp. 793-799, 2000.

[4] A. Nadas, D. Nahamoo and M.A. Picheny, "Speech Recognition Using Noise-Adaptive Prototypes," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 37, No. 10, October 1999.

[5] S.B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. ASSP-28, No. 4, August 1980.

[6] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, Vol. 77, No. 2, February 1989.

[7] H.K. Kim and R.C. Rose, "Cepstrum-Domain Acoustic Freature Compensation Based on Decomposition of Speech and Noise for ASR in Noisy Environments," *IEEE Transactions on Speech and Audio Processing*, Vol. 11, No. 5, September 2003

[8] K. Murphy, "Hidden Markov Model (HMM) Toolbox for Matlab," online at http://www.ai.mit.edu/~murphyk/Software/HMM/hmm.html.