ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

A Factorial HMM Approach to Robust Isolated Digit Recognition in Background Music

Ameya Nitin Deoras and Mark Hasegawa-Johnson

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
deoras@uiuc.edu, jhasegaw@uiuc.edu

Abstract

This paper presents a novel solution to the problem of isolated digit recognition in background music. A Factorial Hidden Markov Model (FHMM) architecture is proposed to accurately model the simultaneous occurrence of two independent processes, such as an utterance of a digit and an excerpt of music. The FHMM is implemented with its equivalent HMM by extending Nadas' MIXMAX algorithm to a mixture of Gaussians PDF. At around 0 dB SNR, the proposed system shows an average relative reduction in word error rate of 57% in the recognition of isolated digits in background music.

1. Introduction

Hidden Markov Models (HMM) have proven to be quite an effective solution to the problem of automatic speech recognition. However, it is widely known that most methods for clean speech recognition fail in the presence of noise, even at moderate signal-to-noise ratios (SNR). Figure 1.1 illustrates the drop in performance of an HMM-based isolated digit recognizer on spoken digits mixed with classical music at different SNRs. The HMM system performs flawlessly down to around 35 dB SNR. Any decrease in SNR beyond that point dramatically increases the word error rate (WER). At 0 dB SNR, i.e. when both speech and background music have equal power, the WER is about 76%. This is the baseline for our later experiments.

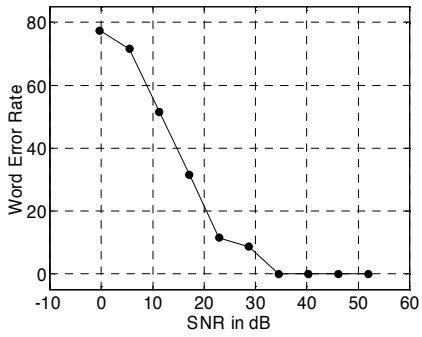

Figure 1.1: The word error rate of an HMM isolated digit recognizer in the recognition of 7 digits in background music at different levels of SNR.

Because of the HMM's sensitivity to noise, a number of techniques, such as spectral subtraction and Weiner filtering, have been developed to make speech recognition robust in noisy environments [1]. However, because most of these techniques assume stationary noise, such as white noise, talking crowds or noisy machines, they cannot be applied to speech in non-stationary noise, such as music or interfering speech, whose statistical properties vary with time.

This paper proposes a Factorial Hidden Markov Model (Factorial HMM or FHMM) approach to solve the problem of recognition of isolated digits in background music. The FHMM simultaneously models both the desired speech and music signals to create an effective speech recognizer in background music, even at low SNR.

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 its component HMMs. 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 isolated digit-in-music recognition system followed by a discussion of its performance in section 5.

2. The Factorial Hidden Markov Model

The class of factorial HMMs was first formalized by Ghahramani and Jordan as an alternative to HMMs [2]. It has been shown that factorial HMMs are better suited to model loosely coupled random processes than HMMs [2], [3]. Efficient algorithms for the estimation of parameters of FHMMs have also been developed [2]. The approach presented in this paper is, however, a little different. The FHMM architecture is used to combine two existing HMMs of independent random processes, such as an utterance of a digit and an excerpt of music. Since the component HMMs have already been trained, no additional training of the FHMM is required. Roweis has shown that an FHMM can be used in such a way to model audio signals from different sources in an auditory scene analysis application [4]. We extend Roweis' method for the recognition of isolated digits in background music.

The motivation behind this model arose from the observed interaction of the log spectra of two signals added in the time domain. Nadas et al. have shown that an additive combination of two sound signals $\bar{Y}(\omega) = \bar{X}(\omega) + \bar{Z}(\omega)$ can be accurately modeled by the element-wise maximum of their log magnitude spectra [5].

$$\log|\bar{Y}(\omega)| \approx \max(\log|\bar{X}(\omega)|, \log|\bar{Z}(\omega)|) \qquad (2.1)$$

This is referred to as the MIXMAX approximation, illustrated in Figure 2.1.

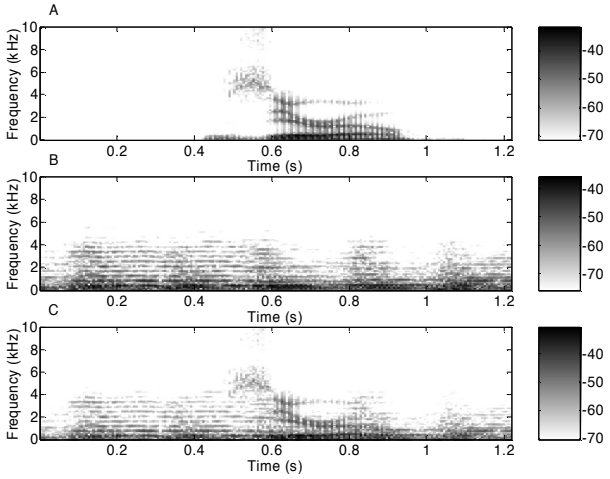

Figure 2.1: Log spectrum MIXMAX approximation; (A) Spectrogram of the digit 'ZERO'; (B) Spectrogram of an excerpt of a string quartet piece; (C) Spectrogram of the combination of speech and music at -0.9 dB SNR. The resultant spectrogram is approximately a piece-wise maximum of the speech and music spectrograms.

2.1. Factorial HMM Architecture

A two-chain factorial HMM, as shown in Figure 2.2, consists of two underlying HMM chains that evolve independently of each other. The output of the FHMM in every frame is the element-wise maximum of the output vectors proposed by each chain independently as expressed in Equation (2.1).

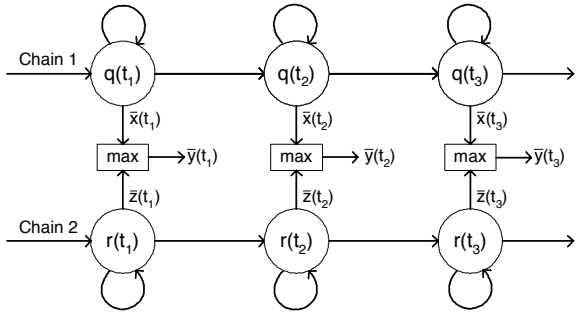

Figure 2.2: A Factorial Hidden Markov Model with two HMM chains with states $q(t)$ and $r(t)$ and outputs $x(t)$ and $z(t)$ respectively. The output $y(t)$ of the FHMM in each frame is the maximum of the outputs proposed by the HMMs.

The inference of the best state trajectory of the FHMM is complicated by the existence of a joint state sequence and a joint PDF. To overcome this difficulty, we implement the FHMM by transforming it into its topologically equivalent HMM. Although this implementation increases the search space (Sec. 2.2), it takes advantage of the MIXMAX approximation (Sec. 2.3) as well as the efficient recognition algorithms that are already available for HMMs.

2.2. Topological Equivalence to an HMM

Consider a factorial HMM, as shown in Figure 2.2, 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 [3]. The transition matrix for the equivalent HMM can be computed in the following manner:

$$a^{FHMM}(i,j \rightarrow k,l) = a^1_{i \rightarrow k} \times a^2_{j \rightarrow l} \quad \begin{matrix} 1 \le i,k \le Q \\ 1 \le j,l \le R \end{matrix} \qquad (2.2)$$

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

2.3. Output Probability Distribution

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 Mel Frequency Spectral Coefficient (MFSC) observation vectors be $\bar{x}_t$ and $\bar{z}_t$ respectively. The output of the FHMM is given by,

$$\bar{y}_t = \max(\bar{x}_t, \bar{z}_t) \qquad (2.3)$$

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

Nadas et al. [5] have shown that, given an expression for the output of a two-HMM system as in Equation (2.3), the PDF of the output can be written as,

$$p_{\bar{y}}(\bar{\lambda}) = p_{\bar{x}}(\bar{\lambda}) F_{\bar{z}}(\bar{\lambda}) + p_{\bar{z}}(\bar{\lambda}) F_{\bar{x}}(\bar{\lambda}) \qquad (2.4)$$

where $F_{\bar{y}}(\bar{\lambda})$ is the CDF of $\bar{y}_t$.

The output probability distribution of each state *(q, r)* of the FHMM can therefore be written as,

$$b_{q,r}(\bar{y}_t) = b^1_q(\bar{y}_t) \int_{-\infty}^{\bar{y}_t} b^2_r(\bar{z}_t) d\bar{z}_t + b^2_r(\bar{y}_t) \int_{-\infty}^{\bar{y}_t} b^1_q(\bar{x}_t) d\bar{x}_t \qquad (2.5)$$

where $b^i_q(\bar{x}_t)$ is the probability of observing vector $\bar{x}_t$ in state q of HMM chain i.

2.4. Extension to a Mixture of Gaussians PDF

Since each HMM state has an output probability density function represented by a mixture of diagonal covariance Gaussians, the output PDF of state q in chain i of the factorial HMM can be written as,

$$b^i_q(\bar{x}_t) = \sum_{m=1}^{M} c^i_{qm} \prod_{p=1}^{n} \left(\frac{1}{\sigma^i_{qm,p}\sqrt{(2\pi)}} e^{-\frac{1}{2}\left(\frac{x_{t,p} - \mu^i_{qm,p}}{\sigma^i_{qm,p}}\right)^2} \right) \qquad (2.6)$$

where M is the number of Gaussians in each mixture, c^i_{qm} is the coefficient for mixture m of state q in chain i, $\bar{x}_t = \left[x_{t,1}, x_{t,2}, ..., x_{t,n}\right]$ and $\left(\sigma^i_{qm,p}\right)^2$ is the element at position *(p,p)* on the diagonal of the covariance matrix Σ^i_{qm}.

Given Equation (2.6), we have shown in [6] that the CDF of the output of each state q of the HMM is given by,

$$\int_{-\infty}^{\bar{y}_t} b_q^i(\bar{x}_t)d\bar{x}_t = \sum_{m=1}^{M} c_{qm}^i \prod_{p=1}^{n} \left(\int_{-\infty}^{y_{t,p}} \frac{1}{\sigma_{qm,p}^i \sqrt{(2\pi)}} e^{-\frac{1}{2}\left(\frac{x_{t,p}-\mu_{qm,p}^i}{\sigma_{qm,p}^i}\right)^2} dx_{t,p} \right) \tag{2.7}$$

The result in Equation (2.7) is an easily implementable formula for calculating the CDF of a mixture of diagonal covariance Gaussians. Therefore, using Equations (2.5), (2.6) and (2.7) the output PDF of each state of the FHMM and its equivalent HMM can be defined.

3. Training Methods

The FHMM recognition system was trained and tested in Matlab with the help of Murphy's Hidden Markov Model Toolbox [7]. Figure 3.1 summarizes the implementation and design procedure followed in the creation of the FHMM system.

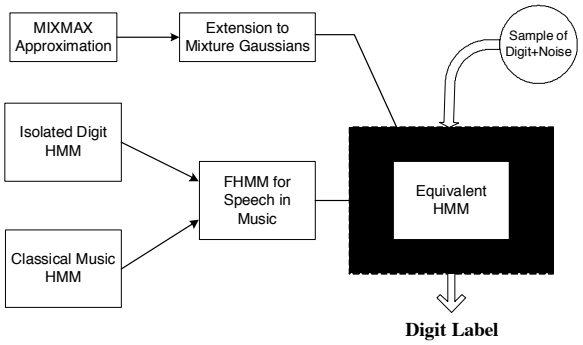

Figure 3.1: Summary of the implementation of the FHMM recognizer for isolated digits in background music.

3.1. Baseline Isolated Digit HMM

The clean-speech isolated digit recognizer was implemented with 10 HMMs, one for each digit from zero to nine. Each HMM was designed with 8 states and 120 Gaussian mixtures per state [8]. In the front-end pre-processor, feature vectors were created by windowing 32ms of data with a Hamming window and computing 20 MFSCs. Consecutive frames were allowed 16ms of overlap. Adjacent frames of MFSC vectors were then concatenated (to create continuity between frames) to produce the observation sequence.

Each digit HMM was trained on 100 utterances of that digit spoken by 50 male speakers from the NIST/TIDIGITS speech corpus [9]. On clean speech, the isolated digit HMM produced a word error rate of 3%.

3.2. Classical Music HMM

It has been shown that HMMs with MFSC feature vectors can be used to accurately model music for genre classification purposes with an accuracy up to 80% [10]. The music model was therefore designed and trained like the isolated digit models, with clips of music used as 'utterances' instead of spoken digits.

To reduce the complexity of the recognition task, the music was restricted to classical music, specifically to string quartets. The classical music HMM was trained on Mozart's Second Divertimento, movement 2, Allegro Molto. The observation sequence was computed by taking consecutive 1.28 s long clips (80 frames) of data from the wave file and generating MFSC observation vectors using the isolated digit front-end. The same front-end was used for training both the isolated digit HMMs and the music HMM because, during recognition, the combination of speech and music are processed with the same front-end processor.

The optimal classical music HMM parameters were determined experimentally. A large number of classical music HMMs were trained by varying the numbers of states, the number of mixture Gaussians per state, the set of allowed transitions (left-right or ergodic) and the size of the training set. The performance of the resulting FHMM (when combined with speech) was then compared for all the models. The optimal model was found to be a left-right HMM with 6 states and 1 Gaussian mixture per state, with a training set containing 40 clips of music.

4. Testing and Results

The following experiments tested the performance of the system in recognizing isolated digits in background music. In each test, an excerpt of music was combined with an utterance of a digit in the time domain, followed by an MFSC observation sequence computation using the same front-end used during training. The system was then presented with the task of determining the digit which when combined with the music HMM produced an FHMM that best modeled the mixed utterance. The baseline was obtained by performing recognition of the noisy speech using just the isolated digit HMMs (see Figure 1.1).

The music data used in the tests was generated by taking 80 frames of data (around 1.28 s), either sequentially or randomly, from a piece of music in wav format. The test pieces were chosen to be the first (Andante) and third (Allegro Assai) movements of Mozart's second Divertimento. The third movement was faster and louder contrasting the slower and quieter first movement. The training and testing data were taken from different movements of the same recording.

Speech data for testing was taken from 5 talkers in the TIDIGITS corpus, none of whom had been used to train the isolated digit HMMs.

The results from each test are presented in Table 4.1. For the task of isolated digit recognition in background music, the FHMM system showed an average relative reduction in WER of 57%.

Table 4.1: Performance of the FHMM system at the task of isolated digit recognition in background music.

Movement	Excerpt and Speaker Sequence	Allowed Digits	Trials	Baseline Performance		FHMM Performance	
				WER	Ave. SNR	WER	Ave. SNR
Allegro Assai	Random	0 - 6	35	83%	-0.39	35%	-0.65
Allegro Assai	Sequential	0 - 6	35	77%	0.55	34%	0.55
Andante	Random	3 - 8	35	70%	4.34	27%	4.77
Andante	Sequential	1 - 8	72	70%	10.43	34%	10.43

5. Discussion

The FHMM system is not as sensitive to changes in SNR as the baseline system. In the first two tests, while the WER of the baseline system decreases by 7% due to an increase of 0.9 dB in SNR, the FHMM WER decreases only by 2.8% (for, in fact, a larger increase in SNR). This is because while the isolated digit HMMs try to identify the speech in spite of the music, the FHMM tries to identify both. Therefore, as long as the MIXMAX approximation holds, the SNR of the combination of speech and music does not significantly affect the performance.

Another feature to notice is the sensitivity of the FHMM to the size of the set of allowed digits. In the fourth test, when the number of digits and utterances is increased by two from the third test, the performance of the FHMM drops more than that of the baseline system. This suggests that the FHMM is more susceptible to confusion from an increased number of possible model combinations.

We have shown similar improvements to the recognition of isolated digits mixed with speech using an identical FHMM framework [6]. The FHMM solution to the problem of recognizing isolated digits in music has a similar range of recognition accuracy at around 0 dB SNR when compared to current highly robust speech recognizers. 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, such as speech or music.

6. Conclusion

We have presented a factorial HMM modeling approach for the recognition of isolated digits in background music.

An isolated digit recognizer was first designed and trained in Matlab for spoken digits at high SNR. Motivated by Petruncio's results from HMM music genre classification using MFSC feature vectors [10], a classical music HMM was also trained with a front-end MFSC observation sequence. The digit and music HMMs were combined using Roweis' Factorial Hidden Markov Model [4] with outputs defined by Nadas' MIXMAX algorithm [5]. The FHMM system was implemented with its equivalent HMM by extending the results from the MIXMAX approximation to a mixture of Gaussians PDF.

At low SNR, the FHMM system was shown to perform isolated digit recognition in background classical music with an average word error rate of 32%, an average relative reduction in WER of 57%.

7. Acknowledgements

This work was supported in part by NSF award number 0132900. Statements in this paper reflect the opinions and conclusions of the authors, and are not endorsed by the NSF.

8. References

[1] H. K. Kim and R. C. Rose, "Cepstrum-Domain Acoustic Feature 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.

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

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

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

[5] 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 1989.

[6] A. N. Deoras and M. Hasegawa-Johnson, "A Factorial HMM Approach to Simultaneous Recognition of Isolated Digits Spoken by Multiple Talkers on One Audio Channel," Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 2004.

[7] K. Murphy, "Hidden Markov Model (HMM) Toolbox for Matlab," May 2003, online at http://www.ai.mit.edu /~murphyk/Software/HMM/hmm.html.

[8] 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.

[9] R. G. Leonard, "A Database for Speaker-Independent Digit Recognition", Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Vol. 4, p 42.11, 1984.

[10] D. Petruncio, "Evaluation of Various Features for Music Genre Classification with Hidden Markov Models," BS Thesis, University of Illinois, March 2002.