SPEECH RECOGNITION USING ADVANCED HMM2 FEATURES

Katrin Weber[1,2], Samy Bengio[1], and Hervé Bourlard[1,2]
[1]IDIAP - Dalle Molle Institute of Perceptual Artificial Intelligence, Martigny, Switzerland
[2]EPFL - Swiss Federal Institute of Technology, Lausanne, Switzerland
email: {weber, bengio, bourlard}@idiap.ch

ABSTRACT

HMM2 is a particular hidden Markov model where state emission probabilities of the temporal (primary) HMM are modeled through (secondary) state-dependent frequency-based HMMs [12]. As shown in [13], a secondary HMM can also be used to extract robust ASR features. Here, we further investigate this novel approach towards using a full HMM2 as feature extractor, working in the spectral domain, and extracting robust formant-like features for standard ASR system. HMM2 performs a non-linear, state-dependent frequency warping, and it is shown that the resulting frequency segmentation actually contains particularly discriminant features. To further improve the HMM2 system, we complement the initial spectral energy vectors with frequency information. Finally, adding temporal information to the HMM2 feature vector yields further improvements. These conclusions are experimentally validated on the Numbers95 database, where word error rates of 15%, using only a 4-dimensional feature vector (3 formant-like parameters and one time index) were obtained.

1. INTRODUCTION

In an attempt to better model the joint temporal/frequency structure of speech, we recently introduced a novel HMM architecture, referred to as HMM2 [12]. HMM2 can be understood as an HMM mixture consisting of a primary HMM, modeling the temporal properties of the speech signal, and a secondary HMM, modeling the frequency properties. A secondary HMM is in fact used at the level of each state of the primary HMM to estimate local emission probabilities of acoustic feature vectors (conventionally done by Gaussian mixture models (GMM) or artificial neural networks (ANN)). Consequently, a conventional (temporal) acoustic feature vector is considered as a fixed length sequence of its components (or subvectors), which has supposedly been generated by the secondary HMM. A similar approach has previously shown some success in computer vision [4, 9, 11].

As described in [14], the HMM2 approach has numerous potential advantages, such as implicit dynamic formant trajectory tracking and automatic spectral warping, possibly permitting easy adaptation to different speakers and conditions. However, HMM2 has not yet shown competitive results in speech recognition, which can be attributed to (1) a restricted modeling power concerning correlations of feature vector components as compared to GMM and (2) a reduced discriminability due to the 'blurring' of important information (such as the positions of spectral peaks) [14]. Here, we introduce a new extension of HMM2, which relies on additional frequency information in the feature vectors, thereby solving the second problem stated above.

On the other hand, it was found that the segmentations obtained by a secondary HMM represent discriminant features for speech recognition, which are related to formant positions [13]. Whereas in this previous work a single secondary HMM was used for feature extraction (basically sharing parameters across all primary HMM states of the HMM2), in the present paper we investigate the use of a full HMM2 system (i.e., with different secondary HMMs for each primary HMM state) in order to extract meaningful structural information such as formant positions, and, as one more new extension, also temporal information (durations and time indices). Fig. 1 shows the resulting system, which is based on two recognition passes: in the first pass, new features are extracted using HMM2, and in the second pass, recognition is performed using a conventional HMM.

In the following, we first describe the formalism of the HMM2 approach and our previous work related to the present paper. Then, we show how the previous HMM2 system can be improved through the introduction of additional frequency information. Thereafter, we explain how meaningful structural information can be extracted using a full HMM2, followed by encouraging experimental results and a brief description of even more promising HMM2 extensions.

2. THE HMM2 APPROACH

Formalism. HMMs are quite powerful statistical models which are used to represent sequential data, e.g. a sequence of acoustic vectors y_1^T in speech recognition. As each acoustic vector y_t can itself be considered as a fixed length sequence of its components $y_t = y_{t,1}^S$, another HMM can be used to model this feature sequence. While a primary HMM models temporal properties of the speech signal, a secondary, state-dependent HMM works along the frequency dimension. The secondary HMM acts as a likelihood estimator for the primary HMM, a function accomplished by GMMs or ANNs in conventional systems. In fact, the state emission distributions of the secondary HMM are modeled

Katrin Weber is currently funded as a Ph.D. student through the Swiss National Science Foundation, project FN 2000-059169.99/1.

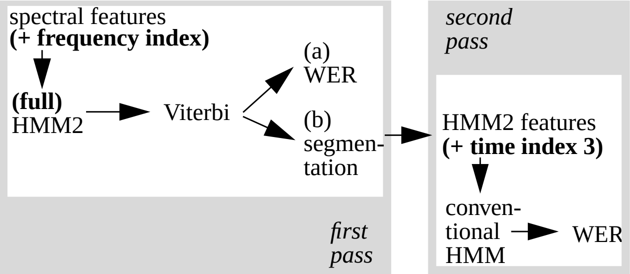


Figure 1: HMM2 system used directly for speech recognition (a), and for features extraction (b). For (b), a second recognition pass, using a conventional HMM, is performed.

by GMMs. Consequently, HMM2 is a generalization of the standard HMM/GMM system (which it includes as a particular case). There are different ways to implement an HMM2 system: (1) using a generalized version of the EM algorithm [1], or (2) realizing HMM2 as one big 'unfolded' HMM and performing conventional EM training [4, 11, 14].

Previous work. Fig. 1 visualizes the two variants of using HMM2: (a) directly for speech recognition and (b) as a feature extractor. In [14], we reported word error rates (WER) of 20.5% (on the Numbers95 database) for the first case. As described above, the secondary HMM acted as likelihood estimator. In [13], we treated a simple variant of the second case, employing a 2-pass system where a single secondary HMM was used as feature extractor. This model was trained on all the training data (regardless of the labeling) and used to extract formant-like structures (in form of the frequency segmentations obtained from the Viterbi algorithm). These were subsequently used as features for standard HMM. We found that (1) the frequency HMM states model indeed spectral regions containing high or low signal energies respectively, and the Viterbi segmentation follows nicely formant-like regions and (2) the segmentation features contain discriminant information, which yield a WER of 37.0% when used as features in a conventional ASR system. Furthermore, when using these segmentation features additionally to noise-robust MFCCs (already including spectral subtraction and cepstral mean subtraction), we observed improved robustness in noisy speech.

3. IMPROVING HMM2

In this paper, we present three important extensions of our previous work (as displayed in boldface in Fig. 1). Firstly, concerning the HMM2 variant using the secondary HMM as likelihood estimator, we show how an additional 'frequency coefficient' is appended to the initial spectral vectors (see section 3.1). The second extension concerns the use of HMM2 as feature extractor: we now use a full HMM2 system (hence with one different secondary HMM for each state of the primary HMM) in order to dynamically extract new HMM2 features (see section 3.2). Finally, these HMM2 features are augmented by temporal information, also extracted with HMM2 during Viterbi decoding (see section 3.3).

3.1 Adding Frequency Information

As conventional HMMs, HMM2 have some difficulties to model duration, which corresponds to frequency bandwidth for the case

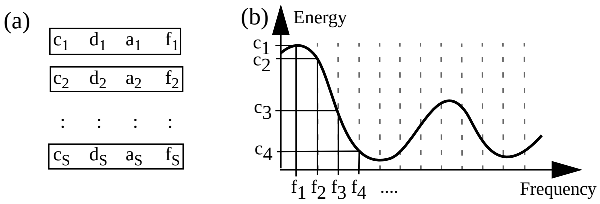
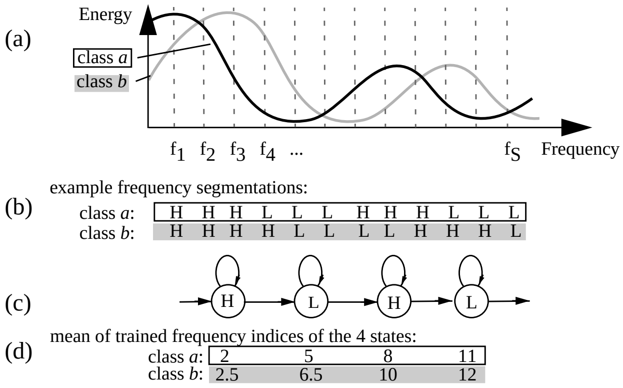


Figure 2: Frequency coefficients: (a) Feature vectors as used in the secondary HMM composed of coefficients c_s, their delta d_s and acceleration coefficients a_s, as well as the frequency coefficient f_s. In (b) it is shown how the 'frequency coefficients' are obtained.

of the secondary HMM. The transition probabilities, which are supposed to model the width of the frequency bands of spectral peaks or valleys, only have a limited influence on the overall likelihood of the sequence. Consequently, a major problem in the application of HMM2 for speech recognition (as identified in [14]) is that the positions of the spectral peaks/valleys in one temporal feature vector do not greatly effect its likelihood. However, as formant positions have shown to be discriminant features for speech recognition [5, 13], it is essential that they are considered in a more sophisticated way in order to obtain a good performance with HMM2. In this paper, we propose a new way to model the frequency positions in a secondary HMM. The idea is to extend each feature vector by its frequency position, as shown in Fig. 2. This has the effect of forcing the Viterbi algorithm to take the frequency position of each feature vector into account during the frequency segmentation.

As an example, let us consider toy speech data of two classes *a* and *b*, both consisting of 2 alternating spectral peaks (H) and valleys (L), resulting in the overall structure HLHL (as shown in Fig. 3) These classes can be distinguished only by the position of the spectral peaks and valleys. Using HMM2 without frequency coefficients, the only way of modeling the differences between *a* and *b* is by the transition probabilities, which, as stated previously, do not have much influence. The two classes are therefore easily confusable. When introducing the frequency coefficients,



example frequency segmentations:

class *a*:	H	H	H	L	L	L	H	H	H	L	L	L
class *b*:	H	H	H	H	L	L	L	L	H	H	H	L

mean of trained frequency indices of the 4 states:

class *a*:	2	5	8	11
class *b*:	2.5	6.5	10	12

Figure 3: Toy speech example: In (a), data assumed to be typical of the classes *a* and *b* are visualized by a black and a gray curve respectively. In (b), an example frequency segmentation is shown for each class. (c) shows a structure of an HMM with alternating H and L states, which is able to model both classes. With an additional trained frequency coefficient (as shown in (d)), discriminability can be ensured.

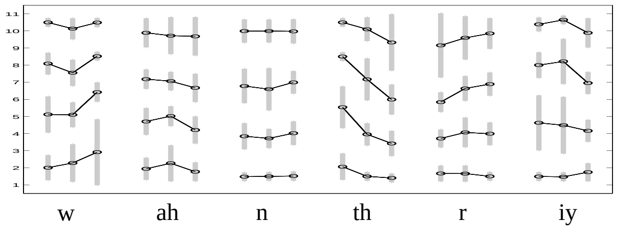


Figure 4: Trained HMM2 parameters for different phonemes. In each column, the means of the frequency indices of the 4 secondary HMM states belonging to the same temporal state are visualized. Vertical bars show the respective variances. The 3 columns belonging to a phoneme correspond to the 3 temporal states.

the Viterbi segmentation of a feature vector is in some way constrained and discriminability will be maintained. In fact, the frequency coefficient is handled in the same way as the other coefficients in a feature vector, i.e. it is modeled by the GMM. The Gaussian mean will correspond to the mean frequency of the modeled frequency band, and the variance should be an indicator for the bandwidth.

While the idea of using an additional frequency coefficient may seem surprising, it is justified in the frequency warping performed by HMM2. As seen later, improved recognition results confirm the suitability of this idea. Naturally, in standard HMMs this frequency coefficient does not give any additional information, as the frequency position of each coefficient is implicitly considered.

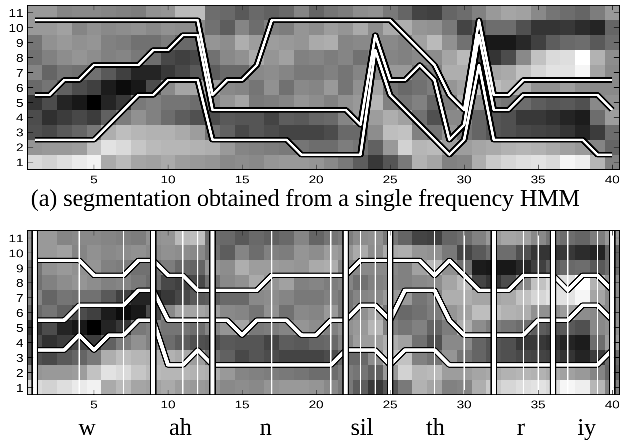
3.2 Using a full HMM2 for feature extraction

HMMs have been used previously to extract structural information such as formant positions from the speech signal [8, 13]. [6] states that the ‘analysis of formants separately from hypotheses about what is being said will always prone to errors’ and that, for a formant analyzer to be optimal, it should be integrated in a recognition scheme. Following the same line of reasoning, we believe that HMM2 offers a suitable framework for extracting speech structures (such as formant positions), which is supported by encouraging experimental results.

As described above, a full HMM2 is used as a feature extractor in a 2-pass recognition system (extension 2). We obtain the temporal and frequency segmentation as a by-product from the Viterbi algorithm performed using HMM2 in the first pass. Contrarily to the case when using a single secondary HMM as feature extractor (as in our previous work), the obtained frequency segmentation depends on the underlying temporal segmentation (i.e. on the hypothesized temporal state sequence of the HMM2). It is obvious that the frequency segmentation may have a different meaning for different temporal HMM states.

3.3 Including a time index

In addition to using the frequency segmentation as features, we can also make use of the temporal segmentation to extract a temporal index and/or a duration parameter (extension 3). This kind of information has successfully been used in speech recognition before, e.g. in ‘trended HMMs’ [3] or in the ‘time index model’ [7]. A duration parameter expresses the total time spent in one speech unit (temporal state or phoneme) according to the HMM2 temporal segmentation (this parameter will be constant for all segmentation feature vectors of the speech unit concerned). A temporal index expresses for each time step the time already spent in the current speech unit.



(a) segmentation obtained from a single frequency HMM

(b) segmentation (and labeling) obtained from a full HMM2

Figure 5: Segmentations obtained (on unseen data) from (a) a single frequency HMM as used previously [13] and (b) a full HMM2 system. In both figures, the horizontal lines correspond to the frequency segmentation. In (b), the vertical lines show the temporal segmentation obtained from the full HMM2 system, where phoneme boundaries are displayed as thick lines, and transitions between temporal states of the same phonemes as thin ones.

4. EXPERIMENTAL RESULTS

Database and HMM2 training. Experiments were carried out on the OGI Numbers95 corpus [2]. 12 frequency filtered filterbank coefficients (including one energy coefficient) [10], additionally normalized, were used as (spectral) features. The 4-dimensional feature vectors consisted of a coefficient, its first and second order time derivatives and its frequency coefficient (here indices from 1 to 12). The HMM2 was realized with HTK [15]. Final models were 80 triphones, each consisting of 3 temporal states. All secondary HMMs had 4 states connected in a looped top-down topology, and an additional non-looped state for the energy. This system was trained globally using the EM algorithm, and Viterbi-based recognition was performed. The trained HMM2 parameters give cues about the structure of the speech signal. In Fig. 4, the Gaussian means of the frequency indices (the 4th component of the feature vectors) are shown for different phonemes of our database (for comparison, those occurring in Fig. 5 were chosen). It can be seen that these parameters vary across phonemes, and that, for a given phoneme, they may also vary in time. Moreover, the corresponding variances are visualized. While the trained means of the frequency index provide information about the position of the frequency bands modeled by the corresponding states, the variances model the respective bandwidths. The figure confirms that some general structural information of the phonemes is modeled. However, the correspondence to formant positions has yet to be thoroughly verified.

In the following, we give experimental results, confirming the utility of the 3 HMM2 extensions described in sections 3 and 3.2.

1. Including frequency information (using HMM2 directly for speech recognition). A WER of 14.0% was obtained, as opposed to 20.5% for a system without frequency index. Although this result is not competitive with state-of-the-art ASR systems (yielding 5.7% on the same database), it is promising as it confirms the validity of the HMM2 model for our purpose: extracting meaningful segmentations along the time and frequency axes.

2. Using a full HMM2 for feature extraction. Fig. 5 shows the obtained temporal and frequency segmentation for an example N95 sentence from an independent test set. In (a), the frequency segmentation obtained with a single secondary HMM is shown. (b) shows the temporal and frequency segmentation obtained by the full HMM2 as described in this paper. To test discriminability of the HMM2 frequency segmentations, they were used as (3-dimensional) features for a conventional ASR system, yielding a WER of 18.6% (which compares to 37.0% for the segmentations of the single secondary HMM).

3. Including a time index. By adding an additional normalized time index to the feature vector (extracted by HMM2 and used in a conventional HMM), the WER could be further reduced to 15.0% (however, the duration parameter was not found to be useful). In our opinion, this a promising result, given the crudeness and the low dimension of the segmentation features.

It has already been shown that segmentation features obtained from a single frequency HMM are quite robust to noise, and that the robustness of a state-of-the-art ASR system can be improved when augmenting noise-robust MFCC feature vector by 3 segmentation features [13]. It is very likely (but still has to be verified) that the segmentation features obtained from the full HMM2, when appended to MFCC feature vectors, shows even greater noise robustness.

features	frequency segmentation from single secondary HMM	frequency segmentation from full HMM2	frequency segmentation and time index from full HMM2
dimension	3	3	4
WER	37.0	18.6	**15.0**

Table 1: Word error rates (WER) on Numbers95 for different segmentation features, using HMM2 as feature extractor.

5. CONCLUSION AND FUTURE WORK

In this paper, we have shown that the HMM2 approach provides us with a sophisticated statistical model, which can be used to extract meaningful structural information such as formant positions from the speech signal. The experimental results, although quite preliminary, confirm the ability of HMM2 for speech modeling in general, and as a robust feature extractor for speech recognition in particular. However, the application of a full HMM2 system as a feature extractor is a completely new research area, which still leaves a lot of space for improvements. For instance, we believe that it is possible to extract even better features with HMM2. For instance, additional representative energy values from the dynamically segmented frequency bands could by added to the segmentation feature vectors. Such an HMM2 can be considered as a dynamic multi-band feature extractor, where the position and width of the frequency bands depend on the temporal segmentation into phonemes and hence varies with the data.

6. REFERENCES

[1] S. Bengio, H. Bourlard, and K. Weber, “An EM Algorithm for HMMs with Emission Distributions Represented by HMMs,” *IDIAP-RR 00-11*, 2000. *ftp://ftp.idiap.ch/pub/reports/2000/rr00-11.ps.gz.*

[2] R. A. Cole, M. Noel, T. Lander, and T. Durham, “New Telephone Speech Corpora at CSLU,” *Proc. Eurospeech*, vol. I, pp. 821-824, Sep. 1995.

[3] L. Deng. A generalized Hidden Markov Model with State-conditioned Trend Functions of Time for the Speech Signal. *Signal Processing*, 27:65-78, 1992.

[4] S. Eickeler, S. Müller, and G. Rigoll, “High Performance Face Recognition Using Pseudo 2D-Hidden Markov Models,” *European Control Conference (ECC)*, Aug. 1999.

[5] P. Garner and W. Holmes, “On the Robust Incorporation of Formant Features into Hidden Markov Models for Automatic Speech Recognition,” *Proc. ICASSP*, 1:1-4, 1998.

[6] W. Holmes, “Segmental HMMs: Modelling Dynamics and Underlying Structure for Automatic Speech Recognition,” *IMA Workshop on Mathematical Foundations of Speech Processing and Recognition*, Sep. 2000.

[7] Y. Konig and N. Morgan, “Modeling Dynamics in Connectionist Speech Recognition - The Time Index Model,” *ICSI TR-94-012*, March 1994.

[8] G. Kopec, “Formant Tracking using Hidden Markov Models and Vector Quantization,” *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. 34, pp. 709-729, Aug. 1986.

[9] S. Kuo and O. Agazzi, “Machine Vision for Keyword Spotting Using Pseudo 2D Hidden Markov Models,” *Proc. ICASSP*, vol. V, pp. 81-84, Apr. 1993.

[10] C. Nadeu, “On the Filter-bank-based Parameterization Front-End for Robust HMM Speech Recognition,” *Proc. Robust’99*, pp. 235-238, May 1999.

[11] F. Samaria, “*Face Recognition Using Hidden Markov Models*,” Ph.D. thesis, Engineering Department, Cambridge University, 1994.

[12] K. Weber, S. Bengio, and H. Bourlard, “HMM2- a novel approach to HMM emission probability estimation,” *Proc. ICSLP, vol. III, pp. 147-150,* Oct. 2000. *ftp://ftp.idiap.ch/pub/reports/2000/rr00-30.ps.gz.*

[13] K. Weber, S. Bengio, and H. Bourlard, “HMM2- Extraction of Formant Structures and their Use for Robust ASR,” to appear in *Proc. Eurospeech*, Sep. 2001. *ftp://ftp.idiap.ch/pub/reports/2000/rr00-42.ps.gz.*

[14] K. Weber, S. Bengio, and H. Bourlard, “A Pragmatic View of the Application of HMM2 for ASR,” *IDIAP-RR 01-23, July 2001. ftp://ftp.idiap.ch/pub/reports/2001/rr01-23.ps.gz.*

[15] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, “*The HTK Book*,” Cambridge University, 1995.