IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse

HMM2- EXTRACTION OF FORMANT STRUCTURES AND THEIR USE FOR ROBUST ASR

Katrin Weber[1,2] Samy Bengio[1] Hervé Bourlard[1,2]

IDIAP-RR 00-42

March 2001

Published in Proc. Eurospeech 2001

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 - 27 - 721 77 11
fax +41 - 27 - 721 77 12
email secretariat@idiap.ch
internet http://www.idiap.ch

1. Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP), Martigny, Switzerland
2. Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland

IDIAP-RR 00-42

HMM2- EXTRACTION OF FORMANT STRUCTURES AND THEIR USE FOR ROBUST ASR

Katrin Weber, Samy Bengio, and Hervé Bourlard

March 2001

Published in Proc. Eurospeech 2001

Abstract: As recently introduced in [1], an HMM2 can be considered as a particular case of an HMM mixture in which the HMM emission probabilities (usually estimated through Gaussian mixtures or an artificial neural network) are modeled by state-dependent, feature-based HMM (referred to as frequency HMM). A general EM training algorithm for such a structure has been developed [2]. Although there are numerous motivations for using such a structure, and many possible ways to exploit it, this paper will mainly focus on one particular instantiation of HMM2 in which the frequency HMM will be used to extract formant structure information, which will then be used as additional acoustic features in a standard Automatic Speech Recognition (ASR) system. While the fact that this architecture is able to automatically extract meaningful formant information is interesting by itself, empirical results also show the robustness of these features to noise, and their potential to enhance state-of-the-art noise-robust HMM-based ASR.

Acknowledgements: This work was partly supported by grant FN 2000-059169.99/1 from the Swiss National Science Foundation. The authors would like to thank S. Pol Font and P. Pujol for their work contributing to the development of frequency filtered features at our institute.

1 INTRODUCTION

State-of-the-art speech recognition systems are based on hidden Markov models (HMM) where the state emission probabilities are typically estimated by Gaussian mixture models (GMM) or artificial neural networks (ANN). As recently introduced, HMM2 consist of standard HMMs where the emission probabilities are estimated by another, state-dependent, 'frequency' HMM [1].

A standard HMM emits a sequence of feature vectors. The estimation of the likelihoods of a feature vector given an HMM state is conventionally based on GMM or ANN. Alternatively, this likelihood can be estimated by a frequency HMM. At each time step, the frequency HMM emits one feature vector in the form of a sequence of its components (usually scalar values). Typically, this feature vector is in the spectral domain, and each of its scalars corresponds to a frequency component. Each state of the frequency HMM is thus described by a one-dimensional probability density function, typically assumed to be Gaussian or a Gaussian mixture. Therefore, the HMM2 parameters are the Gaussian means, variances and mixture weights of the frequency HMM as well as the transition probabilities of the conventional and the frequency HMMs.

Frequency and conventional HMMs can be combined in a recognition system in different ways. A frequency HMM (associated with a certain state of the conventional HMM) estimates the likelihood for a feature vector (which therefore has been decomposed into a sequence of subvectors or scalars). This likelihood is then further processed in the conventional HMM the same way as if it had been estimated by a GMM. In fact, the HMM2 can be 'unfolded' into one big HMM, provided some synchronization constraint is introduced. Training can therefore be done with an EM algorithm as conventionally applied to HMMs. This and similar approaches have been applied before in computer vision [3,4,5] (although the interpretation of the so-called Pseudo2D-HMM differs somewhat from ours). However, in [2] an integrated EM training algorithm which is suited to this particular HMM2 topology (and which therefore avoids the constraints mentioned above) has been proposed.

As we will discuss in more detail in section 2, the HMM2 approach has several advantages compared to state-of-the-art systems, such as the modeling of correlations through the frequency HMM's topology with a parsimonious number of parameters, automatic non-linear frequency warping and dynamic (implicit or explicit) formant trajectory tracking.

In contrast to the HMM2 system introduced in [1], the particular HMM2 system we focus on in this paper uses the frequency HMM to explicitly extract formant structures. The usefulness of formant features for ASR has already been investigated, e.g. in [6,7]. On the other hand, hidden Markov models have successfully been applied to formant tracking before [8]. Here, we propose to use the formant structures extracted by the frequency HMM as features in a conventional HMM system.

In the following, we discuss the motivations of the HMM2 approach. We then demonstrate the abilities of the frequency HMM to extract formant structures, and explain how these can be used for speech recognition. Finally, experimental results showing an improved robustness compared to state-of-the-art noise-robust features are presented.

2 EXTRACTION AND USE OF FORMANT FEATURES

The general motivations of using HMM2 as described above are the following:

- Better feature correlation modeling through the feature-based (frequency) HMM topology. Also, the complexity of this topology and the probability density function associated with each state easily control the number of parameters.

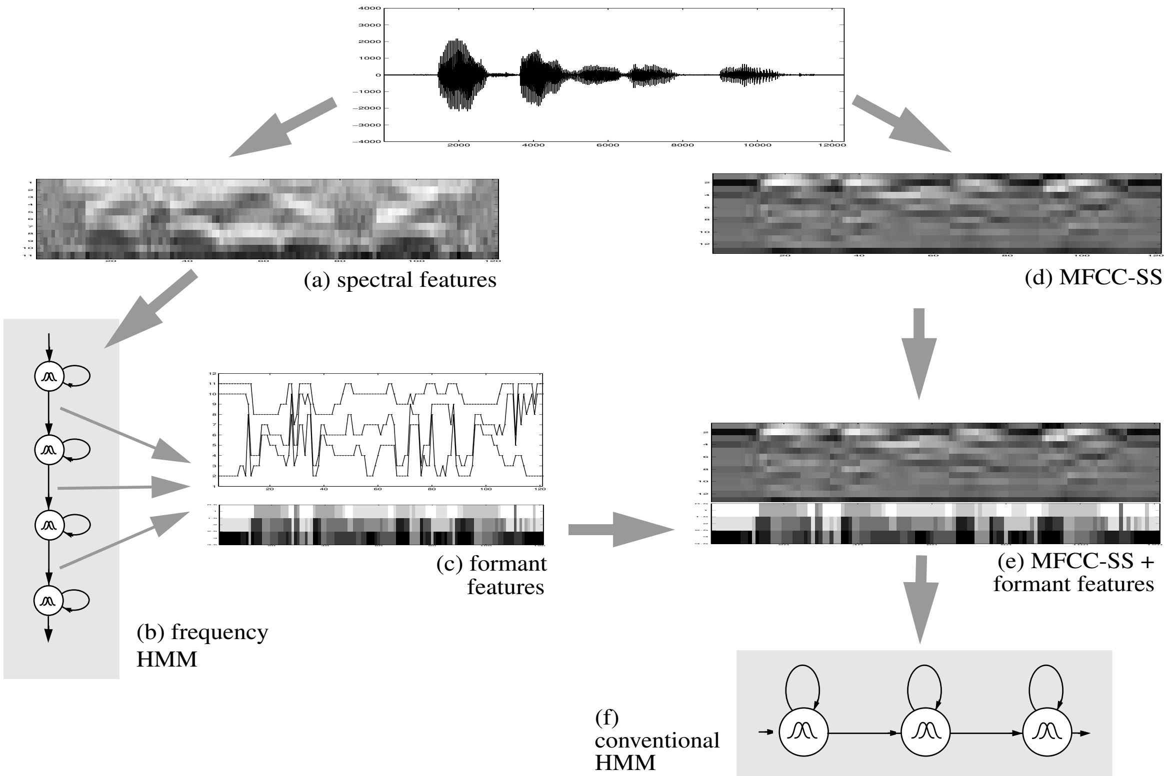

Figure 1: *Extraction of formant features from spectral feature vectors (a) with a frequency HMM (b). Concatenation of the formant features obtained (c) to state-of-the-art noise-robust features (d). The new combined features (e) are then processed in conventional HMMs (f) as usual.*

- Automatic non-linear spectral warping. In the same way the conventional HMM does time warping and time integration, the feature-based HMM performs frequency warping and frequency integration.
- Dynamic formant trajectory modelling. As further discussed below, the HMM2 structure has the potential to extract some relevant formant structure information, which is often considered as important to robust speech recognition.

In the present paper, we mainly focus on the last motivation, as the feature-based HMM is used to dynamically segment the frequency vector into formant-like regions. It is generally agreed that formants are perceptually important features. It is also often acknowledged that spectral peaks (formants) should be more robust to additive noise since the formant regions will generally exhibit a large signal-to-noise ratio. Therefore, the position of these formants in a speech segment could be quite useful for phoneme discrimination.

Figure 1 illustrates how formant-like features can be extracted with a frequency HMM, and then be used as supplementary features in a conventional HMM. Starting from the speech signal, feature vectors in the spectral domain are extracted (a). These features are used to train the frequency HMM (b). After training, the frequency HMM will be used to perform a segmentation along the frequency axis for each spectral feature vector (c). The upper part of (c) shows this segmentation in the time-frequency plane, whereas the lower part visualizes the segmentation features themselves. These 'formant features' are then appended to state-of-the-art noise-robust features (d). Then, the new features (comprising the usual state-of-the-art noise-robust features and the formant features, (e)) are used to train a conventional HMM (f). The usual speech recognition algorithms can then be applied.

As expected, and as further shown later, this approach indeed resulted in the extraction of some meaningful formant information, already quite robust by itself. Further empirical results discussed below also show that complementing standard, already noise-robust, acoustic features with this formant information yielded significant performance improvements in noisy conditions.

3 EXPERIMENTS

The experiments described here were carried out on the OGI Numbers95 corpus [9], which comprises a vocabulary of 30 words (continuously spoken digits). Various noises, e.g. from the Noisex database [10], were added, and different features were extracted.

Our baseline system works with 13 MFCC-SS (i.e., MFCC calculated after spectral subtraction (SS), with cepstral mean subtraction (CMS) and including energy), as well as their first and second order derivatives, making up feature vectors of 39 coefficients. It is important to note that state-of-the-art noise reduction techniques (SS and CMS) have been used in the extraction of these baseline features, so that they are already quite robust to additive as well as convolutional noise. We use the GMM-based HTK system [11] for training and recognition. Final models are 80 context-dependent phonemes with 3 states, each comprising a mixture of 10 Gaussians. On the clean Numbers95 development test set, we obtain a word error rate (WER) of 5.7%.

As features for the frequency HMM we use frequency-filtered filterbanks (FF), as proposed in [12]. These FF features were chosen because they are a rather normalized and decorrelated spectral representation. First and second order temporal derivatives were appended in order to introduce some smoothing. Each of these 33-dimensional temporal feature vectors was rearranged into 11 3-dimensional subvectors, a subvector being composed of an FF coefficient as well was its first and second order time derivatives. At each time step, the frequency HMM emits a 33-dimensional feature vector in the form of a sequence of 11 3-dimensional subvectors.

The topology of the frequency HMM is strictly top-down with loops on each state (see figure 1 (b)). The number of states in the frequency HMM was chosen according to the number of formant regions we aimed to model. For the results reported in this paper, we used 4 states. Training was done with an EM algorithm on all FF data of the Numbers95 train set. Afterwards, the Viterbi algorithm was used to obtain a segmentation (given the trained frequency HMM) for each FF feature vector of the whole database. Such a segmentation typically consists of 3 values indicating the location in frequency of the transitions between the 4 successive frequency HMM states. These values would constitute the new formant feature vector, which can be appended as a second information stream to state-of-the-art features.

3.1 Formant tracking with a frequency HMM

In order to verify that a frequency HMM is indeed able to extract formant structures, we used frequency HMMs trained only on certain parts of the database. The frequency HMM in the right of figure 2 was trained on the vowel ‘ow’. The background of the main part of figure 2 shows a speech segment of FF features (dark and light regions correspond to positive and negative coefficients respectively). In the left part of the image, a segmentation as performed by the HMM on vowel ‘ow’ data is shown. It can be seen that, e.g., state 3 models the high energy region of the FF data. Projected onto the original filterbank spectogram (i.e., before the frequency filtering is applied), the transitions follow quite nicely the maxima (formants) and minima of energy.

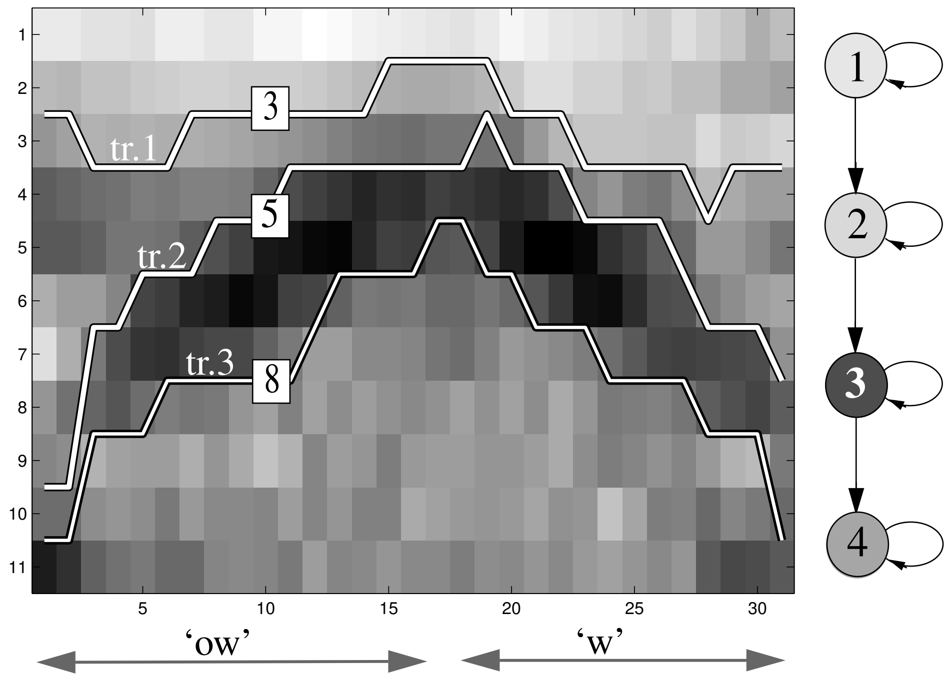

Figure 2: *Segmentations obtained form a frequency HMM trained on the vowel 'ow' for a speech segment containing an example of phoneme 'ow' as in the word 'oh' (left) and of phoneme 'w' as in 'one' (right). The horizontal axis represents time and the vertical axis frequency evolution.*

This conclusion still holds for similar phonemes, e.g. for the vowel 'w' (see right part of the figure), and even for very different phonemes, some formant-like structures are obtained.

Taking a frequency HMM trained on all data and looking at the segmentation results over sequences of several words, we still see formant-like structural information, such as coherent regions in the time/frequency plane modeled by a certain state (an example is displayed in the upper part of figure 1(c)). Although we also find sudden transitions to completely different segmentations (which in fact might partly be due to phonemic transitions), and in case of consonants the interpretation of the segmentation as formant structures no longer holds, this result was encouraging us to go one step further and build a recognition system with the segmentation data as features.

3.2 Recognition experiments

Can the segmentation information obtained from a frequency HMM (as described in the previous section) be useful for speech recognition? To answer this question, we used the segmentation information (obtained from the frequency HMM trained on all data) as features for a conventional HMM, as described previously.

A conventional HMM system was thus trained with feature vectors comprising only 3 components, corresponding to the 3 positions of the segmentation along the frequency axis (e.g., [3 5 8], as shown in Figure 2). Architecture and training procedure of this system are the same as in our baseline system (apart from some parameters due to the different data dimension). We obtain a WER of 43.2% on clean data, what we consider a very good result given the rather crude and low-dimensional features employed here. It shows that our frequency HMM is indeed able to extract meaningful (discriminative) information for recognition.

SNR	39 MFCC-SS (baseline system)	3 formant features	39 MFCC-SS + 3 formant features
clean	5.7	43.2	**5.6**
18	7.4	42.3	**7.3**
12	11.9	49.8	**11.4**
6	23.0	62.2	**21.4**
0	48.6	76.4	**46.6**

Table 1: *WER on Numbers95 with additive Noisex factory noise for different SNR for the noise-robust baseline system (39 MFCCs, extracted after spectral subtraction), a system with 3 formant features only and the combined MFCC-formant-feature system.*

For feature combination with state-of-the-art features, we used the same method as described in [13] and appended the segmental features obtained from the frequency HMM to the MFCCs, thus obtaining feature vectors of 39+3=42 coefficients. Again, we stick to the same system architecture and training procedure as used in our baseline system. For clean speech, we obtain a comparable WER (5.6% vs. 5.7% in the baseline). More results on speech with additive Noisex factory noise can be found in Table 1. While recognition results for high signal-to-noise ratios (SNR) are comparable to those of the baseline system, the word error rate decreases for the case of low SNR. Further tests on car and lynx noise confirmed these results. Overall, we can state with more than 95% confidence that our system works better than the baseline.

3.3 Discussion

The results reported in the previous section are promising, even more so as our present frequency HMM system is still very crude. We believe that with a more sophisticated method for obtaining formant structures, an even better recognition performance can be achieved in noisy speech. One method currently under investigation is to train one distinct frequency HMM for each phoneme, resulting in dynamic, model-dependent features at the level of the conventional HMM. Furthermore, introducing some mechanism in order to smooth the segmentation of the frequency HMM along the time axis could be helpful. Frequency HMMs with different topologies (e.g., number of states) might be investigated, and different spectral features (possibly with a higher frequency resolution) may be used.

4 CONCLUSION

In this paper, we presented a particular instantiation of the HMM2 approach in which formant structures are explicitly extracted by a frequency HMM. These are then used as features in a regular GMM-based HMM systems. We presented experimental results showing that the obtained segmentation might indeed correspond to formant structures, and that it contains some discriminative information which can enhance the robustness of a standard noise-robust ASR system.

5 REFERENCES

[1] 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.*

[2] 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.*

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

[4] 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.

[5] F. Samaria, *"Face Recognition Using Hidden Markov Models,"* Ph.D. thesis, Engineering Department, Cambridge University, Oct. 1994.

[6] P. Garner and W. Holmes, "On the Robust Incorporation of Formant Features into Hidden Markov Models for Automatic Speech Recognition," *Proc. ICASSP*, vol. 1, pp. 1-4, 1998.

[7] L. Welling and H. Ney, "Formant Estimation for Speech Recognition", *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 1, pp. 36-48, 1998.

[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] 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.

[10] A. Varga, H. J. M. Steeneken, M. Tomlinson, and D. Jones, "The Noisex-92 Study on the Effect of Additive Noise on Automatic Speech Recognition," *Tech. Rep.* DRA Speech Research Unit, 1992.

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

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

[13] K. Weber, "Multiple Timescale Feature Combination Towards Robust Speech Recognition," *Proc. Konvens2000/Sprachkommunikation*, pp. 295-299, Oct. 2000.*ftp://ftp.idiap.ch/pub/reports/2000/rr00-29.ps.gz.*