IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse

Increasing Speech Recognition Noise Robustness with HMM2

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

IDIAP-RR 01-36

October 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 01-36

Increasing Speech Recognition Noise Robustness with HMM2

Katrin Weber, Samy Bengio, and Hervé Bourlard

October 2001

Abstract: The purpose of this paper is to investigate the behavior of HMM2 models for the recognition of noisy speech. It has previously been shown that HMM2 is able to model dynamically important structural information inherent in the speech signal, often corresponding to formant positions/tracks. As formant regions are known to be robust in adverse conditions, HMM2 seems particularly promising for improving speech recognition robustness. Here, we review different variants of the HMM2 approach with respect to their application to noise-robust automatic speech recognition. It is shown that HMM2 has the potential to tackle the problem of mismatch between training and testing conditions, and that a multi-stream combination of (already noise-robust) cepstral features and formant-like features (extracted by HMM2) improves the noise robustness of a state-of-the-art automatic speech recognition system.

Acknowledgements: This work was partly supported by grant FN 2000-059169.99/1 from the Swiss National Science Foundation.

1 INTRODUCTION

HMM2 is a particular mixture of hidden Markov models (HMM), where a secondary HMM, working along the frequency dimension of speech, is used to estimate local emission probabilities of a conventional, temporal HMM. The resulting rather flexible model structure has numerous potential advantages, such as a sophisticated modeling of the underlying time/frequency structure of the speech signal and an implicit non-linear frequency warping, leading to systems which may automatically perform formant tracking as well as vocal tract normalization for speaker adaptation.

Recently, considerable progress has been made with HMM2 systems, especially concerning the formant tracking aspect. It has been shown that the HMM2 can segment a speech signal along the frequency axis into high and low energy regions respectively. Therefore, the HMM2 segmentation follows roughly formant-like structures of the speech signal. The fact that formant structures have successfully been used as features for automatic speech recognition (ASR) before ([2],[7]) motivated us to similarly use this HMM2 frequency segmentation as features for speech recognition.

In this paper, we focus on the application of HMM2 to the recognition of speech in noisy conditions. Two variants of using HMM2, namely directly as a decoder for speech recognition and, alternatively, as a feature extractor, are investigated under different conditions. It is demonstrated that HMM2 is in both cases able to outperform conventional HMM systems in the case of heavily degraded signals, given the same (spectral) features. When using HMM2 features in a multi-stream approach to complement noise-robust mel-frequency cepstral coefficients (including spectral subtraction and cepstral mean subtraction, in the following referred to as MFCC-SS), speech recognition results could again be improved significantly.

In the following section, we briefly review the HMM2 approach and its variants, including our previous work. Then, we address the problem of noise robustness, and finally present speech recognition results.

2 HMM2

HMMs are quite powerful statistical models which are used to represent sequential data, e.g. a sequence of acoustic vectors in speech recognition. As each acoustic vector can itself be considered as a fixed length sequence of its components, another HMM can be used to model this feature sequence. In the HMM2 approach, a primary HMM models temporal properties of the speech signal (just as in HMMs conventionally applied to speech recognition), while a secondary, state-dependent HMM works along the frequency dimension. In fact, the secondary HMM acts as a likelihood estimator for the primary HMM, a function accomplished by Gaussian mixture distributions (GMMs) or artificial neural networks in other systems. The state emission distributions of the secondary HMM are then modeled by low-dimensional GMMs. Consequently, HMM2 is a generalization of the standard HMM/GMM system (which it includes as a particular case).

2.1 Motivation

HMM2 provides a very flexible approach to modeling the inherent characteristics of the speech signal. Potential advantages of the HMM2 approach include:

- 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 shown previously [5], the HMM2 structure has the potential to extract some relevant formant structure information, which is often considered as important to robust speech recognition.

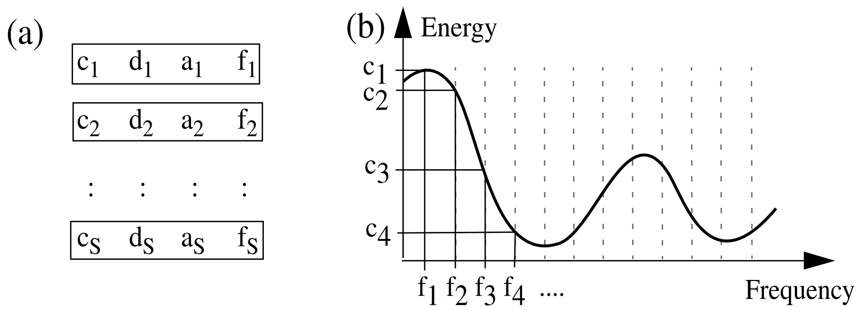

Figure 1: *(a) Feature vectors as used in the secondary HMM composed of coefficients c_s, their delta d_s and acceleration coefficients a_s, as well as the frequency coefficient f_s. In (b) it is shown how the 'frequency coefficients' are obtained.*

2.2 Features for HMM2

For the motivations described in the previous section to hold, it is preferable to use features in the spectral domain for HMM2. This provides us with the physical interpretation of the secondary HMM2 states modeling spectral regions of different energy levels, and permits interpreting the segmentation between these regions as formant-like structures. We here chose normalized frequency filtered filterbank coefficients (FF2, [3]), as they are rather uncorrelated spectral features (apart from correlation near the diagonal, i.e. between coefficients in adjacent frequency bands). Moreover, their performance in conventional HMM system is almost competitive with MFCC-SS in clean speech (however, in additive noise significant degradations were observed for the FF2 features).

A conventional spectral feature vector is split up into a sequence of subvectors, called secondary feature vectors. As illustrated in Fig. 1, a secondary feature vector as used for the HMM2 system is thus composed of an FF2 coefficient (c_s), its first and second order derivatives (d_s and a_s) and a further coefficient reflecting the frequency position of that vector (f_s). Supplementing the 3-dimensional secondary feature vector by such a 'frequency index' has shown significant benefits for speech recognition performance, allowing a better modeling of formant positions (the reader is referred to [6] for more details on the frequency index, its motivations, realization and performance improvements).

2.3 HMM2 variants

In the following, we describe two variants of HMM2 on the application level. Speech recognition with HMM2 is done by the usual Viterbi decoding, and ASR performance can directly be measured by the obtained word error rate (WER). In this case, HMM2 is applied as decoder directly for speech recognition, just in the same way as a conventional HMM, as is visualized in variant (a) of Fig 2.

A by-product of Viterbi decoding (in addition to the sequence of recognized words) is the segmentation. While for conventional HMMs this segmentation is limited to the temporal domain, in the case of HMM2 we obtain an additional segmentation along the frequency dimension, estimated (for each temporal feature vector) from the transitions between the secondary HMM states. Apart from using HMM2 directly for speech recognition, we can use the Viterbi segmentation obtained at each time step as features for a conventional HMM. Furthermore, a temporal index can be calculated from the segmentation between the primary HMM2 states, which has shown to be a beneficial additional component for the new feature vector. Variant (b) of Fig. 2 shows how an HMM2 is only employed for a first recognition pass in a 2-pass system, providing the (temporal and frequency) segmentation features for the second pass. These features are in the following called 'HMM2 fea-

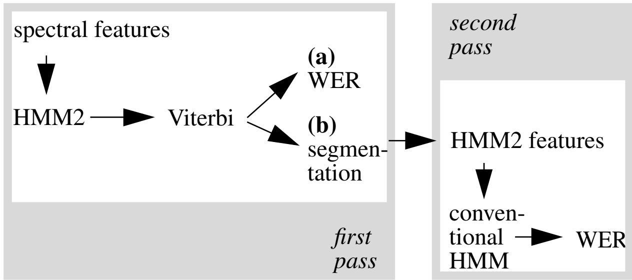

Figure 2: *HMM2 system used directly for speech recognition (a), and for features extraction (b). For (b), a second recognition pass, using a conventional HMM, is performed.*

tures' (Therefore, the term 'HMM2 features' refers to the segmentation features obtained **from** the HMM2, not to be confused with the spectral features used **for** HMM2)

2.4 Building from previous results

Previously, promising results were obtained with both variants of HMM2. In [6], we reported word error rates (WER) of 14.0% (on the clean Numbers95 database, [1]) for variant (a). As described above, here the secondary HMM acted as likelihood estimator. When using the HMM2 in a 2-pass system as feature extractor (variant b), we obtained a WER of 15.0% with only 4-dimensional features on the same task. However, such a full HMM2 system was previously not tested on noisy speech.

In [5], we treated a simplification of variant (b), employing a 2-pass system where a single secondary HMM was used as feature extractor. In fact, the parameters of all secondary HMMs were shared throughout all the primary HMM states. 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 additional features (to complement noise-robust MFCC-SS) for standard HMM, where an improved robustness in noisy speech was observed. In this paper, we will use segmentation features obtained from a full HMM2 as additional features to supplement our noise-robust MFCC. For each time step, the new HMM2 features therefore depend on the present HMM2 primary state (given through the most likely temporal state sequence of the HMM2 feature extractor, given the data), and are therefore class-dependent.

In the following, we will investigate the behavior

- of a **full** HMM2 (as opposed to a simplified version, where the parameters of all secondary HMMs were shared throughout the system, as in [5])
- in **noise** (as opposed to [6], where we investigated a full HMM2, but in clean speech only).

Both variants of applying HMM2 are investigated, and it is shown that:

- HMM2 (when used directly as a decoder for speech recognition) shows a higher robustness to heavily degraded noise, as compared to a conventional HMM, given the same (spectral) features, and that
- HMM2 features (i.e. the structural information extracted from the Viterbi segmentation) provide discriminant information and lead to a significantly improved noise robustness when combining MFCC-SS and HMM2 features in a multi-stream approach.

In the following, we will discuss why HMM2 might be particularly useful for noisy speech, before giving more detailed speech recognition results.

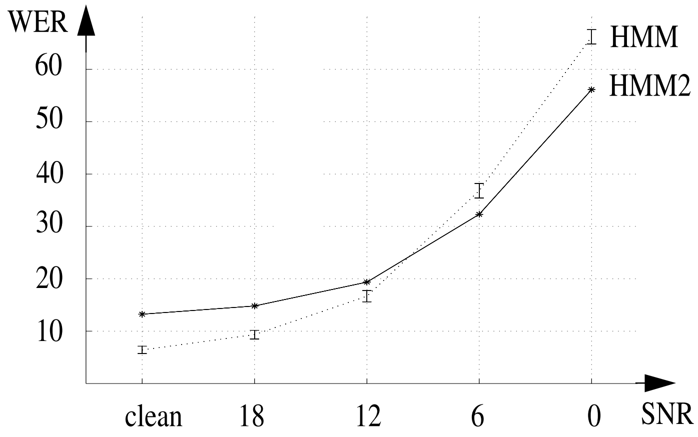

Figure 3: *HMM vs. HMM2 performance for frequency filtered filterbank features, illustrated by the dotted and solid lines respectively. Errorbars for HMM WER show the 95% confidence intervals. The results are for clean speech and car noise at different SNR.*

3 HMM2 AND NOISE ROBUSTNESS

There are several reasons to believe that HMM2 is particularly promising in the presence of noise, for both the straight application of HMM2 as decoder (see the first bullet below) and subsequently the application of HMM2 as a feature extractor (see bullets 1 and 2 below).

- Firstly, it is 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. In many noisy conditions, the overall structure of the speech signal, i.e. the spectogram's partitioning into high- and low energy regions, may largely be unaffected by the noise. As HMM2 relies on these spectral structures, this model may be more tolerant to a large number of distortions.
- Secondly, features extracted from the HMM2 frequency segmentation often correspond to formant-like structures. It is generally agreed that formants are perceptually important features and that they might be robust e.g. against noise and mismatch between training and testing conditions [7]. Moreover, HMM2 formant-like features have already shown good speech recognition performances [6]. If, for the reasons described above, the HMM2 segmentation obtained from the Viterbi algorithm is relatively invariable for different noise conditions given a certain speech unit, and therefore follows the respective formant structures even for highly degraded speech, HMM2 features will show a good robustness to noise.

4 EXPERIMENTS

Experiments were carried out on the OGI Numbers95 corpus [1], corrupted with 3 kinds of additive noises[1] on 4 different signal-to-noise ratios (SNR). 12 FF2 coefficients (including one energy coefficient), additionally normalized, were used as (spectral) features. The 4-dimensional feature vectors consisted of a coefficient, its

1. The noises were partly drawn from the Noisex database [4]. However, the car noise was provided by the IDIAP project partner DaimlerChrysler, which we gratefully acknowledge.

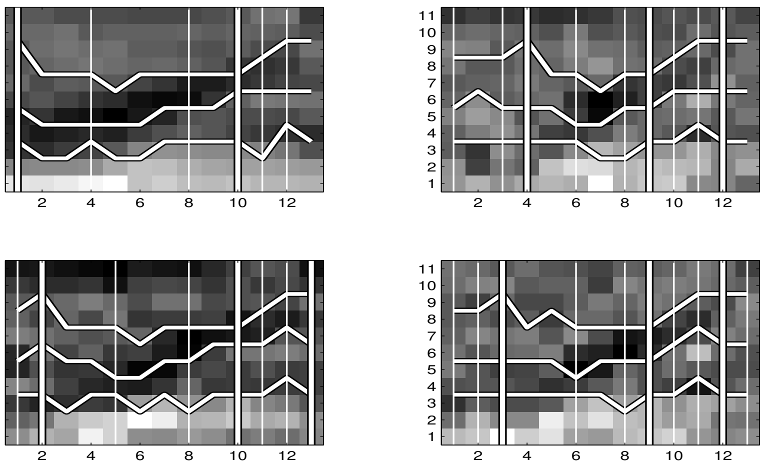

Figure 4: *Temporal and frequency HMM2 segmentation for the same speech segment of the N95 database, for clean speech (upper left) and for speech disturbed with 3 different additive noises at SNR=0. Underlying, the FF2 features are displayed (dark colors correspond to high energy regions). The vertical lines correspond to the temporal, the horizontal ones to the frequency segmentation.*

first and second order time derivatives and its frequency coefficient (here indices from 1 to 12). The HMM2 was realized with HTK [8]. 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, on clean speech only, and Viterbi-based recognition was performed under varying conditions (clean and all noises).

4.1 Results for HMM2 decoder

To realistically compare the performance of the HMM2 system (variant a in Fig. 2) to that of a conventional HMM, we did preliminary tests on both models given the same features (i.e., spectral FF2). Fig. 3 shows results for one noise condition, errorbars indicate the 95% confidence interval. It can be seen that the differences in the performance of these 2 models are statistically significant. While HMM2 is not competitive with conventional HMMs in clean conditions or noisy speech with a high SNR, for heavily degraded noise it easily outperforms the conventional HMMs. In fact, HMM2 is better able to handle the mismatch between training and testing conditions (as training was done on clean speech only). This was confirmed on all other tested noise conditions. Although the obtained results (for both HMM and HMM2 with FF2 features) are not competitive with the state-of-the-art performance (obtained with conventional HMMs, but employing MFCC-SS as features), we feel that this result shows a good potential for applying HMM2 in adverse conditions.

4.2 Results for HMM2 features

In the following, the segmentation features obtained from the HMM2 Viterbi decoding (variant b) are evaluated in noisy conditions. Fig. 4 visualizes the spectograms of FF2 features for the same speech segment (in clean conditions and disturbed with different additive noises), along with the respective HMM2 segmentations. Although the segmentations vary considerably, a general common structure is visible throughout the different conditions. Comparing the HMM2 decoder performance with the recognition rates of HMM2 features, it can be stated that the recognition results obtained from the HMM2 features are slightly inferior to those obtained directly from the HMM2 decoder (but still significantly different from the conventional HMM results shown in Fig. 3) throughout the different testing conditions. In fact, the HMM2 decoder performance seems to be an upper limit for speech recognition using HMM2 features. This confirms our results on clean speech (see section 2.4). However, HMM2 features still have their justification. Firstly, given the crudeness of these features, they perform extremely well (as stated before, for clean speech, we obtain a WER of 15.0%). Secondly, it is straight-forward to combine HMM2 features with noise-robust state-of-the-art features.

SNR	HMM2 features	MFCC-SS	MFCC-SS + HMM2 features
clean	15.0	5.7	**5.7**
18	16.1	6.7	**6.6**
12	20.4	9.3	**9.0**
6	32.8	16.7	**16.1**
0	56.0	35.4	**34.3**

Table 1: *Performance of MFCC-SS and HMM2 features, and their multi-stream combination: WER on Numbers95 at different signal-to-noise ratios: means over 3 different noise types.*

We tested the combination of HMM2 features with MFCC-SS in a multi-stream approach. It has been shown that, while there is a lot of correlation between the 4 dimensions of the HMM2 features themselves, there is not much correlation between the two different feature streams. Furthermore, given the characteristics and different physical interpretation of these two feature streams, it is reasonable to assume that they provide different and supplementary acoustic cues.

Table 1 gives an overview of speech recognition results for HMM2 features, MFCC-SS and their multi-stream combination. In fact, the baseline MFCC-SS speech recognition results were improved for all tested conditions. The obtained results are statistically significantly better than the MFCC-SS only performance (with more than 98% confidence).

As compared to our previous, simplified HMM2 features ([5], described in section 2.4), recognition rates on the HMM2 features have increased by more than 50%, but results on the respective HMM2 features combined with MFCC-SS were not significantly improved. This may indicate that, although by themselves the new HMM2 features perform much better, there is not much more complementary information to the MFCC-SS as already seen in the old and simplified features.

5 CONCLUSION

This paper evaluated two variants of the HMM2 system in noisy speech. Performance improvements were obtained for both HMM2 as decoder as well as feature extractor in heavily degraded noise, as compared to results on conventional HMMs using the same features. However, our HMM2 performance seems still limited by the choice of spectral FF2 features, which cannot compete with robust MFCC-SS in most conditions. Finding more competitive spectral features will be crucial for future HMM2 research. On the positive side, the state-of-the-art MFCC-SS speech recognition results could be improved when supplementing cepstral features with our HMM2 features in a multi-stream approach.

6 REFERENCES

[1] 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.

[2] 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.

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

[4] 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.

[5] K. Weber, S. Bengio, and H. Bourlard, “HMM2- Extraction of Formant Structures and their Use for Robust ASR,” *Proc. Eurospeech*, pp. 607-610, Sep. 2001. *ftp://ftp.idiap.ch/pub/reports/2000/rr00-42.ps.gz.*

[6] K. Weber, S. Bengio, and H. Bourlard, “Speech Recognition using Advanced HMM2 Features,” *To appear in Proc. IEEE ASRU Workshop,* Dec 2001.

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