IDIAP RESEARCH REPORT

EVALUATION OF FORMANT-LIKE FEATURES FOR ASR

Katrin Weber[1,2], Febe de Wet[3],
Bert Cranen[3], Loe Boves[3], Samy Bengio[1], Hervé Bourlard[1,2]

IDIAP-RR 02-04

March 2002

TO APPEAR IN

Int. Conf. on Spoken Language Processing, ICSLP 2002, Denver

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. IDIAP - Dalle Molle Institute of Perceptual Artificial Intelligence, Martigny, Switzerland
2. EPFL - Swiss Federal Institute of Technology, Lausanne, Switzerland
3. Department of Language and Speech, University of Nijmegen, The Netherlands

IDIAP-RR 02-04

EVALUATION OF FORMANT-LIKE FEATURES FOR ASR

Katrin Weber, Febe de Wet, Bert Cranen, Loe Boves, Samy Bengio, and Hervé Bourlard

March 2002

TO APPEAR IN

Int. Conf. on Spoken Language Processing, ICSLP 2002, Denver

Abstract: This paper investigates possibilities to automatically find a low-dimensional, formant-related physical representation of the speech signal, which is suitable for automatic speech recognition (ASR). This aim is motivated by the fact that formants have been shown to be discriminant features for ASR. Combinations of automatically extracted formant-like features and 'conventional', noise- robust, state-of-the-art features (such as MFCCs including spectral subtraction and cepstral mean subtraction) have previously been shown to be more robust in adverse conditions than state-of-the-art features alone. However, it is not clear how these automatically extracted formant-like features behave in comparison with true formants. The purpose of this paper is to investigate two methods to automatically extract formant-like features, and to compare these features to hand-labeled formant tracks as well as to standard MFCCs in terms of their performance on a vowel classification task.

Acknowledgements: We would like to thank Prof. James Hillenbrand for making the AEV database available to us and for his swift reply to all our enquiries. The development of the database was supported by a research grant from the American National Institutes of Health (NIDCD 1-R01-DC01661). Febe de Wet's visit to IDIAP was made possible by the I.B.M. Frye grant. Katrin Weber was funded through the Swiss National Science Foundation, project FN 2000-059169.99/1.

1. INTRODUCTION

Formants are considered to be discriminant features for both human and automatic speech recognition. Increasing research effort is devoted to the use of formant features for ASR, however partly limited by (1) a lack of databases including hand-labeled formants (a consequence of the major effort hand-labeling requires), and (2) the difficulty of automatic methods to reliably estimate formants.

In [2], a database of 12 American English vowels was analyzed. Automatic vowel classification, based on hand-labeled formant values (at some pre-defined times in the vowel) and quadratic discriminant analysis (QDA), was compared to human perception of the speech signal, showing the capacity of true formant values for discrimination.

On the other hand, automatically extracted formant-like features have been used for ASR, showing some potential, especially when combined with state-of-the-art features and where robustness against degradations by noise is required [1,5,7].

Motivated by these findings, the aims of this work are

- to investigate the classification performance of true formant features given state-of-the-art speech recognition methods (namely using HMMs as opposed to the QDA classification performed in [2]). The results obtained can then be seen as an 'optimal' baseline performance for formant-like features;
- to compare the performance of these formant features with that of features conventionally employed in ASR;
- to investigate two methods of automatically extracting features related to formants and evaluate their performance in comparison with the above.

All experiments are based on the 'American English Vowels' (AEV) database [2]. To our knowledge, this is one of only a few databases available that contain hand-labeled formant tracks. We therefore had to restrict our attention to this (for the ASR community) rather small task, only involving vowel classification, rather than speech recognition. While the results presented in the following might not generalize to consonants, they may help to improve the understanding of some of the methods that are available to extract formant-like features. Moreover, they may give an idea about the utility of true and automatically extracted formant features for ASR as compared to other features conventionally employed.

In the following section, we will briefly describe the AEV database, before presenting two methods to automatically extract formant-like features: the 'Robust Formants' method in Section 3 and the 'HMM2 feature extractor' in Section 4. Section 5 then gives an overview of main experimental results.

2. THE AMERICAN ENGLISH VOWELS DATABASE

The speech material that was used in this study is a subset of the database of 'American English Vowels' described in [2]. This database contains recordings of the 12 vowels produced in h-V-d syllables by 45 men, 48 women and 46 children. The speech signals are studio quality and were digitized at 16 kHz. Various acoustic measurements were made for each token in the database, including vowel duration, vowel steady-state times, formant tracks and fundamental frequency tracks. In what follows, the focus will be on the formant tracks, since these values were used as features in our classification experiments.

To obtain the formant tracks, candidate formant peaks were first extracted from the speech data by means of a 14th order LPC analysis. These values were subsequently edited by trained speech pathologists and/or phoneticians. Where unresolvable formant mergers occurred, the higher of the two formant slots affected by the merger was set to zero. The formants corresponding to the leading h's and trailing d's of the h-V-d syllables were not hand-edited.

In [2] it was shown that the vowel classes can be separated reasonably well (in comparison with human performance) by applying a QDA on the values of the first three formants measured at a number of pre-defined times in the vowel.

3. ROBUST FORMANTS

The robust formant (RF) algorithm was initially designed for speech synthesis applications [8]. The algorithm uses the Split Levinson algorithm (SLA) to determine a fixed number of spectral maxima per speech frame. The frequency positions of these maxima are referred to as the 'formants' and they are called 'robust' because the constraints on the algorithm guarantee that the required number of spectral peaks are found under all circumstances. The SLA requires its zero's to lie on the unit circle and to occur in complex conjugate pairs. As a consequence, the resulting number of formant frequencies is always half the LPC order. A major advantage of this method compared to standard LPC analysis in combination with root solving is that the resulting feature tracks are essentially continuous from frame to frame.

The robust formants that were used in this study were calculated every 8 ms over 16 ms hamming windowed segments using a 10th order LPC analysis. The signal was downsampled to 8 kHz and a pre-emphasis factor of 0.98 was applied to the data before windowing. Given the 10th order LPC, 5 'formant frequencies' were calculated per frame. In order to comply with the feature dimension of the hand-labelled formants, 3 of these had to be selected. Various selection criteria were investigated, including methods based on the formants' corresponding bandwidth values. The best classification results (on the training data) were obtained by unconditionally choosing the first, third and fourth formant frequencies calculated by the algorithm.

4. THE HMM2 FEATURE EXTRACTOR

HMM2 [4] is a special mixture of hidden Markov models (HMM), where emission probabilities of a conventional, temporal HMM are estimated by a secondary HMM (one secondary HMM being associated with each state of the temporal HMM), as shown in Figure 1a. While the conventional HMM works along the temporal dimension of speech and emits a time sequence of feature vectors, the secondary HMM works along the frequency dimension (provided that features in the spectral domain are used).

In fact, each temporal feature vector is supposed to be a sequence of its sub-vectors (typically low-dimensional feature vectors, consisting of, e.g., a coefficient, its first and second order time derivatives and an additional frequency index [6]). If a temporal feature vector is supposed to be emitted by a certain temporal HMM state, the associated sequence of (frequency) sub-vectors is in fact emitted by the secondary HMM associated with the current temporal HMM state.

Therefore, the secondary HMMs (in the following also called frequency HMMs) are used to estimate the temporal HMM state likelihoods. In turn, the frequency HMM state likelihoods are estimated by Gaussian mixture models (GMM). As a consequence, HMM2 can be seen as a generalization of conventional HMMs (where higher dimensional GMMs are directly used for state emission probability estimation).

Speech recognition with HMM2 can be done with the Viterbi algorithm, delivering as a by-product the segmentation of the signal in time as well as in frequency. The frequency segmentation of one temporal feature vector reflects its partitioning into frequency bands of similar energy. Supposing that certain frequency HMM states model frequency bands with high energy (i.e., formant-like regions) and others those bands with low energies, the Viterbi frequency segmentation could correspond to formant-like structures.

As features for HMM2, we typically use frequency filtered filterbanks (FF) [3], as they are rather decorrelated features in the spectral domain whose baseline performance is comparable to that of other widely used state-of-the-art features such as mel frequency cepstral coefficients (MFCC). Every 8 ms, a sequence of 12 FF

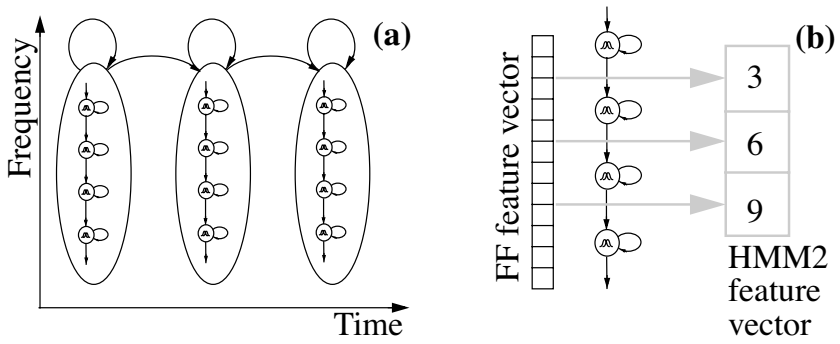

Figure 1: *(a) HMM2 system in the time/frequency plane. The left-right model is the temporal HMM, and in each of its states is a top-down frequency HMM. (b) shows one temporal 'FF' vector (left) as emitted by a frequency HMM. Each of the squares in this feature vector corresponds to a 4-dimensional sub-vector. Grey arrows indicate the frequency positions at which transitions between the different frequency HMM states took place, and the corresponding indices then form an HMM2 feature vector (right).*

coefficients was calculated, which, together with their first and second order time derivatives plus an additional frequency index, form a sequence of 12 4-dimensional sub-vectors.

HMM2 phoneme models consist of three temporal states, and the secondary HMMs associated with them have four frequency states arranged in a top-down topology. Mixtures of 10 4-dimensional Gaussians model the frequency states' emission distributions. For each phoneme in the database, an HMM2 model was trained with the expectation-maximization (EM) algorithm. Subsequently, Viterbi recognition was performed. In the framework of this paper, we did not use HMM2 recognition performance directly, but only the Viterbi segmentation of the signal along the frequency axis, as shown in Figure 1b. For each temporal feature vector, we determined at which point in frequency a transition from one frequency HMM state to the next took place (i.e., between which subvectors (out of the 12) the system passed from frequency HMM states 1 to 2, 2 to 3 and 3 to 4. E.g., in the figure the first HMM2 feature vector coefficient is 3, indicating that the transition from the first to the second frequency HMM state occurred before the third FF vector component). We therefore obtain 3 integer indices (between 2 and 12, corresponding to precise frequency values). These indices can be used as 3-dimensional feature vectors in a conventional HMM.

5. EXPERIMENTS

SET-UP. In this section we compare the classification rates achieved by hand-labeled formants with those obtained by MFCCs, HMM2 features and robust formants. The dataset used was a subset of the AEV database, consisting of the 12 American vowels pronounced by 45 male and 45 female speakers. Where mergers occurred in the hand-labeled formant tracks, we replaced the zeros by the frequency value in the lower formant. To compensate for the absolute differences between the formant frequencies, their values were mel-scaled before they were used in the classification experiments[1]. The mean values that were measured for F1, F2 and F3 are all well below 4 kHz. We therefore decided to downsample the speech data to 8 kHz before feature extraction.

1.Classification experiments were also done using the original, linear frequency values. No significant difference was observed between the tests done on the linear frequency values and the mel-scaled values.

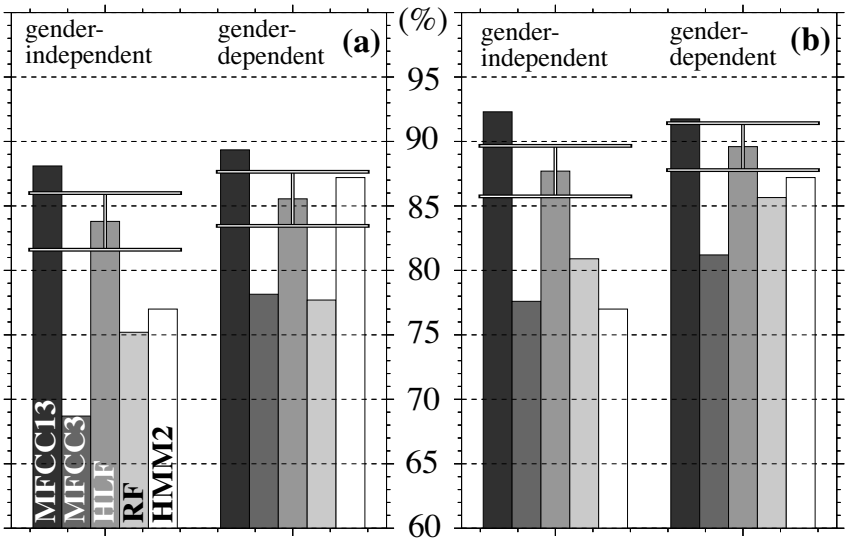

Figure 2: *(a) Average classification rates (%) for gender-independent (left cluster) and gender-dependent (right cluster) models, on MFCC13, MFCC3, HLF, RF and HMM2 features. The 95% confidence intervals given the HLF results are displayed as error bars. (b) Equivalent results on feature sets including first-order time derivatives.*

In comparison with the databases that are typically used in ASR experiments, the AEV database is quite small. Given this limitation, we used a 3-fold cross-validation for the classification experiments. Furthermore, in addition to doing experiments on mixed male and female datasets, we also trained gender-dependent models. For each gender, three independent train/test sets were defined, containing respectively the vowel data of 30/15 speakers. The classification results obtained from the resulting three male and three female models were then averaged to give the gender-dependent results reported below. To train gender-independent models, we used three independent train/test sets, respectively consisting of 60 (30 male, 30 female) and 30 (15 male, 15 female) speakers, and the reported classification rates correspond to the mean classification rate taken over the three outcomes of these experiment.

Five different feature sets were tested:

- MFCC13: 13 mel-frequency cepstral coefficients, including energy, as they are widely employed state-of-the-art ASR features,
- MFCC3: as above, but only using the first three coefficients (and no energy), for comparison since all the other features are also 3-dimensional,
- HLF: hand-labeled formants F1, F2 and F3, as provided with the AEV database,
- RF: robust formants, i.e. automatically extracted formant tracks with the method described in Section 3, and
- HMM2 features, extracted with the method described in Section 4.

For each of these feature sets and for each of the mixed/male/female cross-validation sets defined above, a three state hidden Markov model (HMM) was trained for each vowel using the EM training algorithm implemented in HTK [9]. Each state consisted of a mixture of 10 continuous density Gaussian distributions. Results are shown in Figure 2.

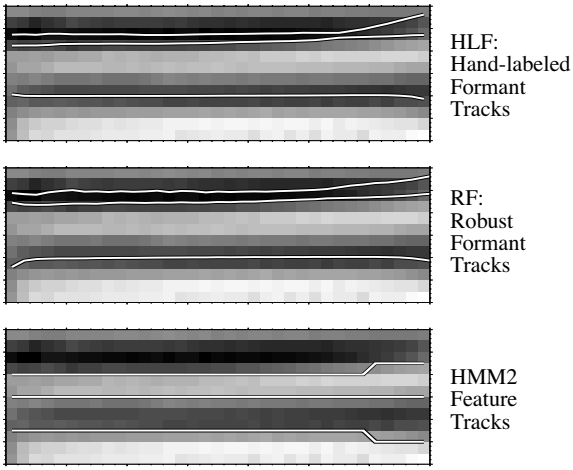

Figure 3: *Tracks of HLF, RF and HMM2 features for one female pronunciation of the vowel in 'heard'.*

RESULTS. The left and right cluster of Figure 2a show classification results on the gender-independent and the gender-dependent models respectively. The different bars in each cluster correspond to the different feature sets. As our goal was to compare performances of HLF to all other features, the 95% confidence intervals were calculated based on the respective HLF results, and are displayed in each cluster as error bars. Figure 2b shows equivalent results, but first-order time derivatives were included in each feature set, therefore doubling the feature dimension. The most important results of these experiments are outlined below.

- Firstly, the HLF features consistently achieved classification rates of over 80%, indicating that they contain discriminant information for vowel classification and that they are suitable to be used as features in combination with state-of-the-art ASR methods (using HMMs, EM training and Viterbi classification).
- Secondly, for the current task, HLF features compare very well with MFCCs. Although MFCC13 features perform significantly better than HLF for all conditions, this is at the price of a much higher feature dimension. MFCCs with the same dimension (MFCC3) perform significantly worse than both MFCC13 and HLF. It should be noted, however, that the choice of using the first three MFCCs might not be the optimal one.
- Thirdly, both automatically extracted formant-like features perform quite well. However, in most cases they are not competitive with their hand-labeled counterparts. An exception are the HMM2 features for the gender-dependent models, which achieve classification results comparable to those obtained by HLFs.

Comparing gender-independent and gender-dependent results (left and right clusters in each sub-figure), it can be seen that generally the gender-dependent systems work better. Comparing Figures 2a and 2b, it can be seen that classification performance was generally improved when including first-order time derivatives in the feature vectors. An exception are the HMM2 features, because their time derivatives are not meaningful due to the HMM2 features' discrete nature and the kind of data present in the AEV database (showing only very little spectral change in each vowel, as can be seen in Figure 3). Therefore, results on HMM2 plus time derivatives have not been included in the figure, and the displayed results correspond to the original 3-dimensional HMM2 features (same as in Figure 2a).

Figure 3 shows feature tracks of HLF, RF and HMM2 features, projected onto a spectogram (based on frequency-filtered filterbank features). The RF tracks are very similar to the HLF. However, HMM2 features are very crude and do not resemble either the HLF or the RF tracks. Nevertheless, the HMM2 method succeeded

in separating high energy from low energy regions, and general trends present in the signal (as the upward tendency for the highest formant at the end of the vowel) are also reflected by the HMM2 tracks.

DISCUSSION. An obvious shortcoming of the present study is that the experiments are restricted to vowel classification, no 'real' speech recognition is done. Unfortunately, this restriction is dictated by the availability of hand-labeled formant data. Even though the current investigation does not give any indications about how well the results will generalize to more typical ASR applications, it does allow for a clear comparison within a restricted domain: formants are well-defined for vowels and, if they are used as features in ASR, they may contribute at least to the separation of the vowel classes.

6. CONCLUSION

In this paper, we have investigated the utility of formant-like features for a vowel classification task. It was confirmed that true (hand-labeled) formants contain discriminant information and can be used as features in state-of-the-art ASR systems. As hand-labeled formants are generally not available as features for ASR, two different methods to automatically extract formant-like features were examined. Both 'robust formants' and 'HMM2 features' achieved interesting classification results, though in most cases inferior to those of the hand-labeled formant features (which in fact can be seen as an upper limit for formant-like feature performance). Even though formant-like features do not equal the higher-dimensional MFCCs in their classification performance, previous studies [1,5,7] have shown that these features can enhance the robustness of MFCCs on real ASR tasks (including consonant classification) in noisy conditions.

7. REFERENCES

[1] 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.

[2] J. Hillenbrand, L.A. Getty, M.J. Clark, and K. Wheeler, "Acoustic characteristics of American English vowels," ***JASA***, vol. 97, no. 5, pp. 3099-3111, May 1995.

[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] K. Weber, S. Bengio, and H. Bourlard, "A Pragmatic View of the Application of HMM2 for ASR," **IDIAP-RR 01-23**, 2001. ***http://www.idiap.ch/~weber.***

[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.

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

[7] F. de Wet, B. Cranen, J. de Veth, and L. Boves, "Comparing acoustic features for robust ASR in fixed and cellular network applications," ***Proc. ICASSP,*** pp. 1415-1418, 2000.

[8] L. F. Willems, "Robust formant analysis," ***IPO Annual report 21***, Eindhoven, The Netherlands, pp. 34-40, 1986.

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