Evaluation of formant-like features for automatic speech recognition[1]

Febe de Wet[a)] Katrin Weber[b,c)] Louis Boves[a)] Bert Cranen[a)] Samy Bengio[b)] Hervé Bourlard[b,c)]

[a)]Department of Language and Speech, University of Nijmegen, The Netherlands
{F.de.Wet, B.Cranen, L.Boves}@let.kun.nl
[b)]IDIAP - Dalle Molle Institute for Perceptual Artificial Intelligence, Martigny, Switzerland
{weber, bengio, bourlard}@idiap.ch
[c)]EPFL - Swiss Federal Institute of Technology, Lausanne, Switzerland

Corresponding author: Febe de Wet

Received:

Suggested running title: Evaluation of formant-like features for ASR
Abbreviated title: Formant-like features for ASR

Abstract

This study investigates possibilities to find a low-dimensional, formant-related physical representation of speech signals, which is suitable for automatic speech recognition. This aim is motivated by the fact that formants are known to be discriminant features for speech recognition. Combinations of automatically extracted formant-like features and state-of-the-art, noise-robust features 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, i.e. robust formants and HMM2 features, and to compare these features to hand-labeled formants as well as to mel-frequency cepstral coefficients in terms of their performance on a vowel classification task. The speech data and hand-labeled formants that were used in this study are a subset of the American English vowels database presented in [Hillenbrand *et al.*, J. Acoust. Soc. Am. **97**, 3099-3111 (1995)]. Classification performance was measured on the original, clean data as well as in (simulated) adverse conditions. In combination with standard automatic speech recognition methods, the classification performance of the robust formant and HMM2 features compare very well to the performance of the hand-labeled formants.

PACS numbers: 43.72.Ne, 43.72.Ar

I Introduction

Human speech signals can be described in many different ways (Flanagan, 1972; Rabiner and Schafer, 1978). Some descriptions are directly related to speech production, while others are more suitable for investigating speech perception. Some descriptive frameworks, of which the formant representation is a well-known example, have successfully been applied to both production and perception.

Speech production is often modeled as an acoustic source feeding into a linear filter (representing the vocal tract) with little or no interaction between the source and the filter. In terms of this model of acoustic speech production, the phonetically relevant properties of speech signals can be characterized by the resonance frequencies of the filter (to be completed with information on the source, in terms of periodicity and power). It is well known that the frequencies of the first two or three formants are sufficient information for the perceptual identification of vowels (Flanagan, 1972; Minifie et al., 1973). The formant representation is attractive because of its parsimonious character: it allows the representation of speech signals with a very small number of parameters. Not surprisingly, many attempts have been made to exploit the parametric formant representation in speech technology applications such as speech synthesis, speech coding and automatic speech recognition (ASR).

A special reason why formants make for an attractive representation of the acoustic characteristics of speech signals is their relation -by virtue of their very definition- to spectral maxima. In the presence of additive noise the lower energy regions of the spectrum of the speech signal will tend to be masked by the noise energy, but the formant regions may stay above the noise level, even if the average signal-to-noise ratio becomes zero or negative (Hunt, 1999). Therefore, one might expect a representation in terms of formant parameters to be robust against additive noise. Automatically extracted formant-like features have shown some potential for noise robustness in automatic speech recognition, especially when combined with state-of-the-art features (Garner and Holmes, 1998; Weber et al., 2001a; de Wet et al., 2000).

Despite its apparent advantages, the formant representation of speech signals has never completely eliminated competing representations. Especially in speech technology there seems to be a strong preference for non-parametric representations of speech signals. These representations are based on estimates of the spectral envelope, if necessary completed by information on the excitation source. Even if the estimate of the spectral envelope is derived from a parametric estimator such as Linear Predictive Coding (LPC) (which can in principle be related to the source-filter model of acoustic speech production (Markel and Gray (Jr.), 1976)), state-of-the-art speech technology systems carefully avoid an explicit interpretation of spectral features in terms of formants.

Given the power of the formant representation in speech production and perception research, its absence in speech technology is disquieting and perhaps undesirable, even if it may not be difficult to "explain" the discrepancy. The single most important disadvantage of the formant representation is that, while resonance frequencies of a linear filter are easy to compute given a small number of characteristic parameters, there is no one-to-one relation between the spectral maxima of an arbitrary speech signal and its representation in terms of formant frequencies and bandwidths. The exact causes of the many-to-many mapping between spectral maxima and formants need not concern us here. What is essential is that despite numerous attempts to build accurate and reliable automatic formant extractors (c.f. (Flanagan, 1972; Rabiner and Schafer, 1978)), there are still no tools available that can automatically extract the "true" formants from the speech in the very large corpora that have become the standard in developing speech technology systems. Labeling of spectral maxima as formants is often only possible if the phonetic label of the sound is known, because there may be more -or fewer- prominent maxima, depending on the spectral characteristics of the source signal, to mention only the most obvious confounding factor. This does not contradict the results of perception studies that suggest that the first three formants are sufficient to identify vowel sounds. The acoustic stimuli used in those experiments are almost invariably constructed so as to avoid spectral maxima related to the excitation signal.

The many-to-many relation between spectral maxima and formants is not the only reason why speech technology systems avoid formant representations. Not all speech sounds are equally well suited to be described in terms of formant frequencies in the sense of resonance frequencies of a linear filter. Nasals and fricatives, for example, can only be accurately described if anti-resonances are specified in addition to the resonances. It is well known that anti-resonances can mask formants to the extent that they no longer appear as spectral maxima. This masking can even occur in vowels that are nasalized because of their phonetic context. Last but not least, the voice source may contain spectral peaks and valleys, which may also affect the spectral peaks in the radiated speech signal. Thus, even if it were possible to accurately and reliably label spectral maxima as formants, one would still be faced with the fact that many portions of the speech signals that must be processed show fewer (or more) spectral maxima than the number predicted by acoustic phonetic theory. Most of the search algorithms that are used in ASR algorithms are designed to deal with feature vectors of a fixed length. Recently, attempts have been made to design ASR systems that are able to cope with missing data (Cooke et al., 2001; de Veth et al., 2001; Renevey and Drygajlo, 2000; Ramakrishnan, 2000), but still in the context of search algorithms that require fixed-length feature vectors. In these approaches "unreliable" parameter values obtain a special treatment in the computation of the distance between a feature vector and the models of the speech sounds that have previously been trained. However, none of these systems use formants as features. One of the few recent ASR systems that do try to use formants (in addition to non-parametric spectral features) is (Holmes et al., 1997). In (Holmes et al., 1997) it is proposed to overcome the problem of labeling spectral maxima as formants by introducing a confidence measure on the formant values. The approach proved to be quite successful, but only for a limited task and a small data set.

Most modern ASR systems rely on very large labeled corpora to train probabilistic models. Due to the lack of tools to compute formants reliably and accurately, experts are needed to add formant labels to the speech. This makes it very difficult to provide sufficiently large training corpora for the development of formant-based processing. Yet, the theoretical attractiveness of formant representations has motivated several attempts to overcome this hurdle. This paper extends this line of research by investigating two techniques to extract formant-like features that may overcome at least one of the problems in more conventional formant extraction techniques. The methods we investigate, i.e. two-dimensional hidden Markov models (HMM2) (Weber et al., 2000) and Robust Formant extraction (RF) (Willems, 1986), can be guaranteed to find a fixed number of "formants" in each spectral slice. The details of these techniques will be explained in Sections III and IV. By guaranteeing to deliver a fixed number of formant-like features for each frame, these techniques avoid problems in the search of the ASR engine that would arise if the number of parameters were allowed to vary from frame to frame. The research in this paper is focused on automatic speech recognition. Therefore, we will not make references to applications of the techniques in speech synthesis and speech coding in the remainder of this paper, despite the fact that the RF technique was developed in that context.

There is an obvious area of tension between the definition of "true formants" in terms of resonances of the vocal tract on the one hand, and a formant extraction technique that guarantees to deliver a fixed number of formant-like features for each frame of a speech signal on the other. It is unlikely that what these automatic techniques deliver always corresponds to vocal tract resonances, even if the parameters can be proven to relate to spectral maxima. This raises the question whether the formant-like features delivered by these automatic extraction techniques are as powerful as the true formants that could have been measured by expert phoneticians when it comes to identifying speech sounds.

In order to compare the classification performance of (true) formants measured by phoneticians and (imperfect) formant-like features extracted by means of HMM2 and RF, a speech corpus with hand-labeled formants is required. Such corpora are extremely rare, because - as was explained above - their construction requires an enormous amount of time and expertise. One of the few corpora that does include hand-labeled formants is the *American English Vowels* (AEV) database presented in (Hillenbrand et al., 1995). The details of the AEV corpus are described in Section II. Here it is sufficient to say that the corpus consists of 12 American-English vowels, pronounced in /h-V-d/ context by 45 men, 48 women and 46 children. The identification of all vowel tokens was checked in perception experiments.

Despite the large effort spent in generating the AEV corpus, its size is very small by ASR standards, and the corpus only contains information about vowels. Consequently, promising results obtained with the AEV corpus may not generalize to continuous speech that will inevitably contain consonants, both voiced and voiceless. However, the goal of the research reported in this paper was not to develop a full-fledged alternative automatic speech recognizer. Rather, we aim at a better understanding of the contribution that formant-like representations of speech can make to the improvement of automatic speech recognition. More specifically, the aims of the research reported here are

- to investigate whether the classification performance of (true) formants measured by phoneticians represents an upper limit for the performance of (imperfect) formant-like features extracted by means of HMM2 and RF. This will be done for two different classification techniques, i.e.

    1. Discriminant Analysis, where we used straightforward Linear Discriminant Analysis (LDA) instead of Quadratic Discriminant Analysis (QDA) that was used in the original AEV paper (Hillenbrand et al., 1995);
    2. Hidden Markov Models (HMMs), which are considered state-of-the-art in today's ASR.

- to interpret the classification performance of automatically extracted formant-like features in terms of their resemblance to true formants. This should improve our understanding of the importance of the relation between vocal tract parameters in speech production and acoustic features for automatic speech recognition.

- to investigate the claim that formant-like features are inherently robust against additive noise, because they relate to spectral maxima that will stay above the local spectral level of additive noise. For practical reasons, this part of the study is limited to automatically extracted formant-like features.

The rest of this paper is organized as follows: Section II gives an overview of the protocol according to which the AEV database was created. The RF algorithm is the subject of Section III and the HMM2 feature extractor is described in Section IV. Section V reports on the experimental set-up and the results of the classification experiments. The results are followed by a discussion and conclusions in Sections VI and VII.

II Database of American English Vowels

The speech material that was used in this study is a subset of the database of American English vowels (AEV) described in (Hillenbrand et al., 1995). This section provides some information on the construction of the database and the labeling of the formant data. Interested readers are referred to the original paper for a complete overview of the database.

Amongst other things, the AEV database contains recordings of the 12 vowels (/i, ɪ, ɛ, æ, ɑ, ɔ, ʊ, u, ʌ, ɝ, e, o/) produced in /h-V-d/ syllables by 45 men, 48 women and 46 children. The /h-V-d/ syllables were produced in isolation, not within a carrier phrase. Full details on the screening and selection of the subjects can be found in (Hillenbrand et al., 1995). During the recordings, the subjects read from one of 12 different randomizations of a list containing the key words corresponding to the /h-V-d/ syllables. They were given as much time as needed to practice the task and to demonstrate their ability to pronounce the key words correctly. On average, three recordings were made per subject. Unless there were problems with recording fidelity or background noise, the tokens from the subject's first reading of the list were taken up in the database.

The recordings are all studio quality and were digitized at 16 kHz with 12 bits amplitude resolution. Various acoustic measurements were made for each token in the database, including vowel duration, vowel steady-state times[2], 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 14^{th} order LPC analysis. These values were subsequently edited by trained speech pathologists, phoneticians, or both. In addition to the LPC peaks overlaid on a gray-scale spectrogram, labelers were also provided with individual LPC or Fourier slices where necessary. The labelers were allowed to repeat the LPC analysis with different parameters and to hand edit the formant tracks. The formant tracks were only hand edited between the start and end times of the vowels, i.e. the formants corresponding to the leading /h/ and trailing /d/ of the /h-V-d/ syllables were not manually labeled.

Where irresolvable formant mergers occurred, zeros were written into the higher of the two formant slots affected by the merger. Irresolvable mergers occurred in about 4% of the data. F1, F2, and F3 were measured for all the signals, except for utterances that contained irresolvable mergers. F4 tracks were only measured if they were clearly visible in the peaks of the LPC spectrum. In 15.6% of the utterances F4 could not be measured. We therefore decided to limit the scope of the formant feature set to the first three formants.

Given that the mean values that were measured for F1, F2, and F3 were all well below 4 kHz, we decided to downsample the speech data to 8 kHz for our own experiments. All acoustic analyses adhered to the same time resolution used in (Hillenbrand et al., 1995). Specifically, all analyses used a frame rate of one frame per 8 ms. This allows a frame-to-frame comparison of the hand-labeled formants with the formant-like features generated by the two automatic extraction techniques.

III Robust Formants

The robust formant (RF) algorithm was initially designed for speech coding and synthesis applications (Willems, 1986). The algorithm uses the split Levinson algorithm (SLA) to determine a fixed number of spectral maxima for each speech frame. Instead of directly applying a root solving procedure to a standard LPC polynomial to obtain the frequency positions of the spectral maxima, a so-called singular predictor polynomial is constructed from which the zeros are determined in an iterative procedure. All the zeros of this singular predictor polynomial lie on the unit circle, with the result that the number of maxima that are found is guaranteed to be half the LPC order under all circumstances. The maxima that are located in this manner are referred to as the "formants" found by the RF algorithm.

After the frequency positions of the RF formants have been established, their corresponding bandwidths are chosen from a pre-defined table such that the resulting all-pole filter minimizes the error between the predicted data and the input. The frequencies at which the zeros of the singular predictor polynomial occur are close to the frequencies at which the zeros of the classical root solving procedure occur, as long as these are close to the unit circle (i.e. as long as the true formants have small bandwidth values). This property ensures that the most important formants are properly represented.

For our goal (as was the case for speech coding and synthesis), the RF algorithm has two major advantages over standard root solving of the LPC polynomial (or searching for maxima in the spectral envelope derived from the LPC coefficients). First, the SLA guarantees to find a fixed number of complex poles -corresponding to "formants"- for each speech frame. This helps to avoid labeling errors (e.g. F3 labeled as F2) since there are no missing formants. In addition, the algorithm tends to distribute the complex poles uniformly along the unit circle. Consequently, the formant tracks are guaranteed to be fairly smooth and continuous (as one would expect the vocal tract resonances to be). A potential disadvantage of the SLA is that it cannot handle formant mergers in a way that resembles the procedure used in (Hillenbrand et al., 1995). Because of the tendency of the SLA to distribute poles uniformly along the unit circle, formant mergers are likely to result in one or two "resonances" that are shifted away (in frequency) from the true resonances of the vocal tract.

As was mentioned in Section II, the AEV data was downsampled to 8 kHz. It is usually assumed that there are four vocal tract resonances in this frequency band. However, the data in (Hillenbrand et al., 1995) show that F4 could not be found in 15.6% of the vowels. The scope of this study is therefore limited to F1, F2, and F3. Moreover, in the AEV database the mean value (taken over all the relevant data) of F4 is 3.536 kHz ($\sigma = 135.5$) for males and 4.159 kHz ($\sigma = 174.7$) for females. Thus, it is clear that an automatic formant extraction procedure applied to the AEV corpus must be able to deal with a potential discrepancy between the "true" number of formants in the signal and the requirement that only the first three formants must be returned.

For the RF extractor, the simplest way to cope with the requirement that only three formants should be found is to use a 6^{th} order LPC analysis[3]. However, the accuracy of the LPC analysis is bound to suffer if a 6^{th} order analysis is used to analyze spectra with four maxima. In these cases an 8^{th} order LPC would seem more appropriate, although it would introduce the need to select three RFs from the set of four.

Given these constraints, there are a number of possible choices that can be made concerning the calculation of the RFs. We considered two of these: (1) calculate three RF features per frame (RF3); (2) calculate four RF features per frame and use only the first three (3RF4). These two sets of RF features were subsequently calculated every 8 ms over 16 ms Hamming windowed segments. The output of the two procedures was evaluated by means of a frame-to-frame comparison with the hand-labeled formants. The mean Mahalanobis distance between the resulting RF3 and 3RF4 features and the corresponding hand-labeled formants (HLF) are given in Table I.

Table I about here.

The results in Table I show that the RF features are closer to the HLF features if the order of the analysis is chosen according to the gender-specific properties of the true formants. If there is a mismatch between the number of spectral peaks the algorithm tries to model and the number of spectral maxima that actually occur in the data, the distance between the automatically derived data and the hand-labeled data increases. Thus, the distance between the RFs and the hand-labeled formants decreases if the order of the analysis corresponds to the inherent signal structure. In the rest of this paper we will present results for both gender-dependent and gender-independent data sets. Because the RF3 features yielded the smallest Mahalanobis distance for the mixed data set, these will be used in the gender-independent experiments. In the gender-dependent experiments, the RF3 and 3RF4 features will be used for the female and male data, respectively.

IV The HMM2 Feature Extractor

In this section, we introduce the most important characteristics of the HMM2 approach. HMM2 is a special mixture of hidden Markov models (HMM), in which the emission probabilities of a conventional, temporal HMM are estimated by a secondary HMM (Weber et al., 2001b). As shown in Figure 1, one secondary HMM is associated with each state of the temporal HMM. 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, and emits a frequency sequence of feature vectors, provided that features in the spectral domain are used.

In fact, each temporal feature vector can be seen as a sequence of sub-vectors. The sub-vectors are typically low-dimensional feature vectors, consisting of, for example, a coefficient, its first and second order time derivatives and an additional frequency index (Weber et al., 2001c). If such a temporal feature vector is to be emitted by a specific temporal HMM state, the associated sequence of frequency sub-vectors is emitted by the secondary HMM associated with the corresponding 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.

Figure 1 about here.

Frequency filtered filterbanks (FF) (Nadeu, 1999) are typically used as features for HMM2, because they are decorrelated in the spectral domain. In many ASR tasks the baseline performance of the FF coefficients has been shown to be comparable to that of other widely used state-of-the-art features such as mel frequency cepstral coefficients (MFCCs). For the HMM2 systems that were used in this study, a sequence of 12 FF coefficients was calculated every 8 ms, which, together with their first and second order time derivatives plus an additional frequency index, form a sequence of 12 4-dimensional sub-vectors. Each square in the vector labeled "FF feature vector" in Figure 1 therefore represents a 4-dimensional sub-vector.

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 be interpreted as an alternative way to represent formant-like structures.

For each temporal feature vector, we determined at which point in frequency (i.e. between which sub-vectors) a transition from one frequency HMM state to the next took place. For example, in Figure 1 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 sub-vector. In the case of 4 frequency HMM states connected in a top-down topology (as seen in Figure 1), we therefore obtain 3 integer indices (corresponding to precise frequency values). In our classification experiments, these indices were used as 3-dimensional feature vectors in a conventional HMM.

A HMM2 design options

The design of an HMM2 system can vary substantially, depending, for example, on the task and on the data to model. There are a number of design option which determine the performance of an HMM2 system. These include issues like *model topology* (which needs to be considered both in the time and the frequency dimension), the addition of *frequency coefficients*, different *initialization* possibilities as well as different (combinations of) segmentation strategies that can be applied for *training and test* purposes. In the following, each of these issues is shortly discussed.

As a first step in HMM2 design, a suitable *topology*, i.e. the number and connectivity of the temporal and the frequency HMM states, has to be defined. In this study, we chose a strict "left-right" (without any state skipping) topology for the temporal HMM (such as typically used for HMMs used in ASR) and an equivalent "top-down" topology for the frequency HMM. It should be noted, however, that the choice of topology is by no means limited to these options: e.g. the frequency HMM can also have an ergodic, a tree- or trellis-like, or any other topology (Weber et al., 2000).

Given the restriction of a left-right/top-down HMM2 topology, the number of HMM states of the temporal and the frequency HMMs can still be varied. However, in all experiments described in this paper, the frequency HMM had 4 states. This choice was motivated by the task at hand (i.e. extracting three formant-like features from each speech frame), as well as the characteristics of the data used. Different numbers of states for the temporal HMM were tested. In the first instance, a very simple HMM2 feature extractor was realized using just one HMM2 model, which had one temporal state with four frequency states, and which was trained on all the training data, independent of the class labeling. Obviously, such a model cannot be used directly for speech recognition. Nevertheless, a forced alignment of the data given this model delivers a frequency segmentation of each temporal data vector and therefore "HMM2 feature vectors". These features should - in a very crude way - represent frequency regions of similar energy.

Furthermore, 12 phoneme-dependent HMM2s with a similar topology (i.e., one temporal HMM state) were tested, as well as 12 phoneme-dependent HMM2s with 3 temporal states. In both cases, a 4-state frequency HMM was associated with each temporal state. These HMM2 models were trained with the expectation maximization (EM) algorithm, and Viterbi recognition was subsequently performed. Both of these systems can be applied directly as a decoder for speech recognition, or, as in the context of this paper, for feature extraction. Although the quality of phone-dependent HMM2 feature extraction suffers from the fact that HMM2 recognition is error-prone, using such a system (as opposed to, e.g. using just one HMM2 model) is motivated by the assumption that the "... analysis of formants separately from hypotheses about what is being said will always be prone to errors" (Holmes, 2000). In fact, it can be confirmed that, in terms of recognition rates, the features obtained from the phone-dependent HMM2 systems generally perform better than those obtained from a single model.

A further HMM2 design decision concerns the use of a *frequency coefficient* as an additional component of the frequency sub-vectors. It has been shown that this frequency information improves discrimination between the different phonemes (Weber et al., 2001c). However, the impact of the frequency coefficient is different depending on whether it is treated (1) as an additional feature component (feature combination) or (2) as a second feature stream (likelihood combination). Moreover, in the latter case, additional parameters are required, i.e. the stream weights.

The *initialization* of the HMM2 models can be done in different ways. For instance, assuming a linear segmentation along the frequency axis, the initial features can be chosen such that an equal number of sub-vectors is assigned to each of the 4 frequency states. Alternatively, as formant frequencies are provided with the AEV database, these can be used to obtain an initial non-linear frequency segmentation. Another option is to assume an alternation of spectral valleys (L) and spectral peaks (H), i.e. assigning values to the frequency states which force an HLHL or LHLH segmentation along the frequency axis.

HMM2 feature vectors can be obtained in two different ways, depending on whether or not the labeling is known. For the *training* data, we typically know the phoneme labeling of all the speech segments. Therefore, forced alignment can be used to align these speech data to the corresponding HMM2 model and extract the segmentation. Alternatively for the training data, and imperatively for the *test* data, a real recognition using all phoneme-dependent HMM2 models can be used. The segmentation finally extracted by the HMM2 system corresponds to the segmentation produced by the HMM2 phoneme model which has the highest probability of emitting the given data sequence. Obviously, the HMM2 system makes recognition errors, resulting in sub-optimal HMM2 feature vectors, i.e. feature vectors extracted by the "wrong" HMM2 phoneme model.

In this study, all of the design, initialization and training/test options introduced above, as well as combinations of them, were tested. However, it is beyond the scope of this paper to give an exhaustive overview of these results. The models that were used to obtain the results reported on in Section V all had a 3-state, left-right topology in the time domain and a 4-state top-down topology in the frequency domain. Frequency coefficients were not used as a second feature stream but were included as additional feature components in the frequency sub-vectors. The gender-independent HMM2 models were initialized with an LHLH segmentation while the gender-dependent models were initialized according to the hand-labeled formant frequencies' segmentation. The HMM2 features that were used for training were obtained by means of forced alignment while those that were used for testing were obtained from a free recognition. Training and testing were done with HTK (Young et al., 1997) and the HMM2 systems were realized as a large, unfolded HMM, which is possible when introducing synchronization constraints (Weber et al., 2001b).

Finally, it should be pointed out that results from a previous study have shown that adding first order time derivatives does not improve the classification performance of HMM2 features (Weber et al., 2002). In that study, it was argued that this result can be attributed to (1) the nature of the AEV data, exhibiting only very few spectral changes (see Section V.D for a graphical illustration), in conjunction with (2) the very crude nature of the HMM2 features. Often, the frequency segmentation of one phoneme would be the same for all time steps, thus the time derivatives are zero. In other cases, oscillations between two neighboring segmentations were observed, which give equally meaningless derivatives.

V Experiments and Results

In this section, we describe the design and execution of the experiments that were performed on the AEV database in order to investigate the classification performance of two sets of automatically extracted formant-like features. The behavior of the RF and HMM2 features is compared to the results obtained using the hand-labeled formants that are included in the AEV database.

In section A, the overall design of the experiments is described. Section B reports on the results of classification experiments based on Linear Discriminant Analysis (LDA). These experiments enable us to relate our results to those reported in the original paper on the AEV database (Hillenbrand et al., 1995). In section C, the results of classification experiments based on HMMs are presented. These experiments are included to investigate whether the proven classification performance of hand-labeled formants with LDA generalizes to the classification performance obtained with the EM procedures that are dominant in the ASR community.

To strengthen the link with current research in automatic speech recognition, all classification experiments were repeated with acoustic features that are used in most conventional ASR systems, i.e. MFCCs, which describe the spectral envelope in a small number of essentially orthogonal coefficients. Usually, 10 to 15 MFCCs are needed to obtain a sufficiently accurate description of the spectrum. In our experiments, two sets of MFCCs were used. The first set comprises 12 coefficients to account for the spectral envelope and one energy feature. Since this set contains more than four times as many independent coefficients as the representation in terms of F1, F2 and F3 we also used a subset consisting of c_1, c_2, and c_3, i.e., the first three MFCCs that are related to the shape of the spectrum.

In order to explain some of the classification results, we also present a number of graphical illustrations of the differences and similarities between hand-labeled formant values and the RF and HMM2 features in Section D. Finally, Section E reports on the classification performance of the automatically extracted formant-like features in (simulated) noisy acoustic conditions.

A Experimental set-up

In all the experiments reported on in this section, a subset of the AEV database was used, i.e. the 12 vowels (/i, ɪ, ɛ, æ, ɑ, ɔ, ʊ, u, ʌ, ɝ, e, o/) pronounced by 45 male and 45 female speakers. Only the vowel part of these utterances were taken into consideration, because the formant tracks of the leading /h/s and trailing /d/s were not hand-edited. Where mergers occurred in the hand-labeled formant tracks (c.f. Section II), the zeros were replaced by the frequency values in the lower formant slot, i.e. two equal values were used. This procedure allowed us to treat all vowels in the same way, including those where mergers occurred. Alternatively, we might have replaced the merged formants with frequencies slightly below and above the value that is given in the AEV database, but it is unlikely that this would have affected the results.

In keeping with what has become standard practice in ASR, the formant frequencies were mel-scaled before they were used in the classification experiments[4]. In comparison with the databases that are typically used in ASR experiments, the AEV database is quite small. Given this limitation, a 3-fold cross-validation was used for the classification experiments. The classifiers (LDA and HMM) were trained on two subsets of the data, and tested on the third one. Thus, each experiment consisted of a number of independent tests. Moreover, all tests were performed in two conditions, i.e. *gender-independent* and *gender-dependent*. The gender-independent data sets were defined as three non-overlapping train/test sets, each containing the vowel data of 60(train)/30(test) speakers, with an equal number of males and females in each set. For the gender-dependent data, three independent train/test sets were defined for males and females, respectively. Each train/test set consisted of 30(train)/15(test) speakers. For the gender-independent data sets, the classification results reported below correspond to the mean value of the three independent tests. The gender-dependent results were obtained by averaging the classification results of six independent experiments (three male and three female).

Five different feature sets are relevant to the experiments in this section:

- HLF: hand-labeled formants F1, F2, and F3, as provided with the AEV database;
- RF: robust formants, formant tracks extracted automatically using the method described in Section III;
- HMM2: HMM2 features, extracted according to the method described in Section IV;
- MFCC13: 12 mel-frequency cepstral coefficients, together with an energy measure (c_0 in this case) as an example of commonly-used, state-of-the-art ASR features[5];
- MFCC3: as above, but using only three coefficients (c_1, c_2, c_3) for comparison, since all the other feature sets are 3-dimensional.

B LDA results

In (Hillenbrand et al., 1995), a number of discriminant analyses were performed in order to determine how well the vowel classes could be separated based on the different acoustic measurements. A quadratic discriminant analysis (QDA) was applied in a leave-1-out jack-knifing procedure and all the male, female and children's data (except for the vowels /e/ and /o/[6]) were used. Using the linear frequency values of F1, F2, and F3 measured (within one frame) at steady state (stst), 81.0% of the vowels could be correctly classified. The corresponding formant values measured at 20% and 80% vowel duration (20%80%) yielded 91.6% correct classification. A combination of the three values (20%stst80%) resulted in a classification rate of 91.8%. Human classification for the same data (based on the complete /h-V-d/ utterances) was 95.4% correct. These values indicate that the vowel classes can be separated reasonably well (in comparison with human performance) by the steady state values of their first three formants. Information about patterns of spectral change clearly enhances the distinction between classes.

This section reports on a similar (but not identical) experiment in which the LDA classification performance of the RF, HMM2 and MFCC features was compared to the classification rate achieved by the HLF features. An LDA was used instead of a QDA, all frequency values were mel-weighted and only the male and female data were taken into consideration. The training and test data were divided according to the 3-fold cross-validation scheme described in Section A. The feature values were all measured at the same time instants in the vowel as for the experiments described in (Hillenbrand et al., 1995). The results for the gender-independent data are given in Table II and those for the gender-dependent data in Table III. As our goal was to compare the performance of the HLF features with that of the other features, the 95% confidence intervals corresponding to the HLF results are indicated in brackets.

Tables II and III about here.

With the exception of the steady state results, the classification rates achieved by the HLF features are in good agreement with the corresponding values reported in (Hillenbrand et al., 1995). The difference observed for the steady state results can probably be attributed to the difference between the QDA used in (Hillenbrand et al., 1995) and the LDA used in the current study.

The values in Tables II and III show that, with the exception of the MFCC13 features, the HLF features outperform all the other features in terms of vowel classification rate. The difference between HLF and the other results is much larger for the gender-independent experiments than for the gender-dependent experiments. This observation suggests that, in the gender-independent condition, three hand-labeled formant frequencies represent more information on the identity of the vowel classes in the AEV set than three RF, HMM2 or MFCC features. This is not surprising, since the formant features incorporate substantial know-how from expert phoneticians and speech pathologists. If an essential part of that prior knowledge, i.e. the gender of the speakers, is given to the other feature extractors, their performance is substantially enhanced. For instance, in the gender-independent experiments the classification rate achieved by the RF features is clearly inferior to the HLFs' performance. The corresponding difference in classification performance is much smaller in the gender-dependent experiments.

The classification performance of the HMM2 features is substantially lower than the results obtained for the other feature sets. Obviously, the vowel classes are not linearly separable given these features at just one, two or three different instances in time. While the HMM2 features at any given moment may not be sufficient to discriminate between the vowel classes, the additional information required to do so may be provided by a complete temporal sequence of HMM2 features. This presupposition will be investigated in the following section within the framework of HMM recognition.

The MFCC13 features achieve classification rates which compare very well with those of the HLF features. Although they perform slightly better than the HLF features in the gender-dependent experiments, this difference is not significant. This result indicates that, for the current vowel classification task using LDA, three HLF features and 13 MFCCs are equally able to discriminate between the vowel classes.

The MFCC3 features do not seem to provide a description of the vowel spectra that is able to compete with HLF or RF features in terms of vowel classification. However, it should be kept in mind that choosing the first 3 MFCCs as features is probably not the best choice we could have made. In a control experiment we used Wilk's lambda to rank the MFCCs in terms of explained variance. This resulted in different feature combinations for different experimental conditions. However, the set that was most frequently observed (for the gender-dependent data) was c_2, c_4, and c_5. Using these 3 MFCCs instead of c_1, c_2, and c_3 improved the gender-dependent classification rates by about 2% (on average). Although this is a substantial improvement, it does indicate that, in combination with LDA, more than 3 MFCC features are required to compete with HLF and RF features on a vowel classification task.

Classification performance is determined by two factors, i.e. the degree of noise in the features and the overlap between the vowels in the feature space. The data in Tables II and III show that all the feature types that were evaluated in this experiment generally yield much better results for the gender-dependent data sets. This observation may be explained by the fact that the vowel classes are better separated in a gender-dependent feature space. However, the RF and HMM2 features clearly benefit more from the gender separation than the HLF and MFCC features. This seems to suggest that, for the RF and HMM2 features, the gender separation also achieved a certain degree of noise reduction in the features themselves. For instance, according to the Mahalanobis distance measures in Table I, the gender-dependent RF features approximate the HLF features much better than their gender-independent counterparts. For the HMM2 features the biggest advantage of the gender separation (in terms of reducing the noise in the features) is probably the fact that the original classification of the vowels (during the HMM2 feature extraction process) improved.

C HMM classification rates on clean data

The classification rates in Tables II and III were obtained by means of an LDA. In discriminative training algorithms such as LDA, the aim of the optimization function is to achieve maximum class separability by finding optimal decision surfaces between the data of the different classes. However, the recognition engines of most state-of-the-art ASR systems are trained using a Maximum Likelihood (ML) optimization criterion. The training algorithms therefore learn the distribution of the data without paying particular attention to the boundaries between the different data classes. Although discriminative training procedures have been developed for ASR, they are not as commonly used as their more straightforward ML counterparts. The LDA classification described in the previous section also required a time-domain segmentation of the data. In real-world applications this kind of information will not be available. The aim of the next experiment is therefore to evaluate the classification performance of the different feature sets using HMMs that were derived by means of ML training.

Towards this aim, we compared the vowel classification rates achieved by the five different feature sets introduced in Section A. With the exception of the HMM2 features, the first order time derivatives of all the features were also included in the acoustic feature vectors. In a previous study (Weber et al., 2002), it was shown that adding temporal derivatives to the HMM2 features does not improve performance, most probably due to the very crude quantization of these features, which causes most of the time derivatives to become zero. The resulting feature vector dimensions for the HLF, RF, HMM2, MFCC13, and MFCC3 features were therefore 6, 6, 3, 26 and 6.

Classification experiments were conducted using both the gender-independent and the gender-dependent data sets defined in Section A. For each of the vowels in the AEV database and for each acoustic feature/data set combination, a three state HMM was trained. The EM algorithm implemented in HTK was used for the ML training (Young et al., 1997). Each HMM state consisted of a mixture of 10 continuous density Gaussian distributions. The results are shown in Table IV. The values in the last column of Table IV correspond to the dimensions of the different feature sets. Once again, the 95% confidence intervals corresponding to the HLF results are indicated in brackets.

Table IV about here.

According to the results in Table IV, the HLF features consistently achieved classification rates of almost 90% correct. Even though these values are significantly lower than those measured in the LDA experiments, they do indicate that, in principle, the HLF features are suitable to be used as features in combination with state-of-the-art ASR methods, i.e. using HMMs, ML training and Viterbi classification. However, in practical applications the use of hand-labeled features is not really feasible.

A remarkable difference between the LDA and HMM experiments is the difference in the classification rates achieved by the HMM2 features: these features perform much better in combination with HMMs than LDA. Table IV shows that, for the gender-dependent data, the HMM2 features not only outperform the MFCC3s but also approximate the performance of the HLF and RF features, in spite of their lower feature dimension.

The data in Table IV also show that, for the current vowel classification task, HLF features compare very well with MFCCs. Although the MFCC13 features outperform their HLF counterparts on both gender-independent and gender-dependent data, 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. Once again, the choice to use the first 3 MFCCs is probably not optimal. In order to be completely fair towards the MFCCs, 3 coefficients should have been selected by means of, e.g. principle component analysis.

Comparing gender-independent and gender-dependent results, it can be seen that, in general, the gender-dependent systems work better, even in the case of HLF features. This observation is in good agreement with the results of the LDA experiments. Another similarity between the HMM and LDA results is the fact that the classification performance of the automatically extracted formant-like features are especially gender-dependent. As was argued before, the large improvement of the performance of the RF and HMM2 features in the gender-dependent condition is most probably due to the combination of the fact that there is less noise in the raw data (because of the gender specific measurement techniques) and, again, removal of gender-related overlap between feature values. Although not to the same extent as the formant-like features, the performance of the MFCC3 features is also enhanced by incorporating gender-information. Only the performance of the MFCC13 features seems to be insensitive to gender differences. This may be due to the capability of the EM training algorithm to capture the difference between female and male spectra in the 10 Gaussians in each state. The larger number of parameters in the MFCC13 feature space is also likely to have improved the recognition performance.

D Graphical examples

In this section we will illustrate, by means of a graphical example, the differences and similarities between the hand-labeled formants and the corresponding RF and HMM2 features for the vowel /ɝ/. Figure 2 shows feature tracks of HLF, RF and HMM2 features, projected onto two different "spectrograms". In both instances the y-axis corresponds to frequency index, the x-axis to time and darker shades of gray to higher energy levels. The spectrogram in Figure 2(a) corresponds to the mel-weighted log-energy within each frame. The mel-scaled filterbank that was used to scale the energy values consisted of 14 filters that were linearly spaced in the mel frequency domain between 0 and 2146 mel (0 and 4000 Hz). The spectrogram in Figure 2(b) was derived from the corresponding FF features that were used to train the HMM2 feature extractor.

Figure 2 about here.

The data in Figure 2 show that the RF feature tracks are fairly similar to the HLFs. However, it should be kept in mind that this is a gender-dependent example. The Mahalanobis distances in Table I indicated that the differences between the HLF and RF tracks are substantially larger in the gender-independent data, where wrongly labeled spectral peaks are more frequent. The example suggests that the spectrum of this vowel contains multiple peaks in the F2-F3 region, and that the human labeler has consistently preferred a peak at a lower frequency than the automatic RF procedure. In addition, the RF features exhibit more frame-to-frame variance than their hand-labeled counterparts - especially for F3. The dip in the F3 track at the vowel onset may be due to the fact that there the lower frequency peak preferred by the human labelers throughout was so strong that the RF procedure could find it, despite its close proximity to F2. So far, we have not been able to verify whether this type of frame-to-frame variation is related to those parts of the vowels where the human labelers had most problems in finding the "correct" spectral peaks. Neither is it clear whether this variation has affected the classification performance of the RF features, relative to the more smooth HLF features. From an articulatory point of view the smooth HLF feature tracks seem to be more plausible than the slightly more "noisy" RF features. However, it may be that the RF features are a better descriptor of the acoustic signal than the manually smoothed HLFs. This observation raises the question whether it is at all possible for a tractable automatic procedure to emulate the expert knowledge that is implicitly encoded in the HLF features. Fortunately, from the point of view of automatic classification an exact emulation is not essential: if avoidable measurement noise in the RF features is indeed avoided, their performance is equivalent to the HLF features.

The HMM2 features are very crude and do not resemble either the HLF or the RF tracks. The crudeness is due to the fact that the HMM2 features are derived from 12 FF features, instead of spectral envelopes sampled at multiple equidistant frequencies. Moreover, due to their very nature (they indicate transitions between regions of low and high spectral energy, rather than spectral peaks) the HMM2 tracks can at best approximate the shape of true formant tracks, not their position on the frequency axis. However, the feature tracks in Figure 2(b) indicate that, in the FF domain, the HMM2 method succeeded in separating high energy from low energy regions. General trends present in the signal (such as the upward tendency for the highest formant at the end of the vowel) are also reflected by the HMM2 tracks. As was noted before, the HMM2 features' time derivatives are not meaningful because of their discrete nature and the kind of data present in the AEV database (showing very little spectral change in each vowel).

E HMM classification rates on noisy data

In this experiment, the MFCC13, RF and HMM2 models that were used for the experiments described in Section C were tested in noise. The models were trained only on clean data. Noisy acoustic conditions were simulated by artificially adding babble and factory noise to the test data at SNRs of 18, 12, 6, and 0 dB. The babble and factory noise were both taken from the Noisex CD (Noisex, 1990). For obvious reasons the HLF features could not be included in this experiment.

Figure 3 gives an overview of the classification performance of gender-dependent models tested in noise. Classification rate is shown as a function of SNR for both babble and factory noise. Similar, but slightly inferior, results were obtained for the gender-independent models. (These results are not shown here.)

Figure 3 about here.

In Section I it was argued that, in the presence of additive noise, the lower energy regions in speech spectra will tend to be masked by the noise energy, but that the formant regions/spectral maxima may stay above the noise level, even if the average signal-to-noise ratio becomes zero or negative. This line of reasoning gave rise to the hypothesis that a representation in terms of formants or formant-like features should be inherently robust against additive noise. However, the results in Figure 3 do not support this hypothesis. In fact, the figure shows that the recognition performance of all three systems deteriorates in noise. While the performance of the different features is comparable at SNRs of 18 dB and higher, the MFCC13 features clearly outperform the formant-like features at lower SNRs. To a certain extent, this result may be explained by the fact that the MFCC13 system has a total of 26 features at its disposal, while the dimensionality of the RF and HMM2 systems is restricted to 6 and 3 features, respectively. The higher order acoustic feature vectors - which may contain redundant information in clean conditions - seem to be better at maintaining system performance in adverse acoustic conditions.

For all three systems the drop in recognition rate is more severe in factory noise than in babble noise. Factory noise also seems to affect the RF features more than HMM2. The type of performance degradation shown in Figure 3 is equivalent to results obtained for other databases in comparable simulated noise conditions (e.g., (de Wet et al., 2000)).

In principle, the argument that spectral maxima may stay above the noise level seems to be plausible. However, the RF features - which are supposed to model spectral maxima - clearly fail in noisy acoustic conditions. This observation suggests that the RF algorithm is "misled" by the information between the spectral peaks, such that it is no longer capable to find the maxima that should still be in the spectra. This limitation can be overcome by an algorithm which is capable of finding spectral maxima without being hindered by the misleading information between the peaks. Such an algorithm was recently proposed in (Andringa, 2002).

The failure of the HMM2 system at low SNRs may be explained as follows: for heavily degraded speech, the number of recognition errors made by the HMM recognizer embedded in the feature extractor is bound to increase. As a result, the corresponding HMM2 features will be calculated by the "wrong" HMM2 feature extractor, i.e. the HMM2 model corresponding to the wrong phoneme will give the best likelihood score and will therefore be chosen for feature extraction. Recognition errors made by the HMM2 feature extractor and the conventional HMM recognizer (which uses the erroneous HMM2 features) accumulate, which will forcibly lead to severe degradations at low SNRs.

VI Discussion

The research reported on in this paper intended to investigate the contributions that formant representations of speech signals can make to automatic speech recognition, in clean and especially in noisy acoustic conditions. Since the design of the experiments required the availability of reliably hand-labeled formants, the extent of this study is limited to the AEV database. This database is not representative of "normal" speech, if only because of the fact that the phonetic contexts of the vowels are limited to /h-V-d/. In a sense, therefore, the AEV database constitutes a best case platform for research on the added value of formant(-like) features. Within this context, it was confirmed that hand-labeled formants are suitable features for vowel classification, both in combination with discriminant analysis and state-of-the-art ASR systems. However, given the fact that hand-labeled formants cannot be used in practical situations, two different methods to extract formant-like features automatically were examined, i.e. RF and HMM2.

The results reported in Sections V.B and V.C showed that, for clean data (with the exception of HMM2 features in combination with LDA), the classification performance of both these formant-like feature sets compares very well to the performance of hand-labeled formant features. RF features consistently outperformed HMM2 features, most probably due to the fact that the HMM2 features are very coarsely quantized. Moreover, the HMM2 features are only 3-dimensional, whereas the RF features have additional delta's and therefore 6 dimensions. This observation shows that hand-labeled formants are certainly not the only parsimonious representation of the spectral envelope that enables accurate vowel classification. Representations that yield a regular and consistent description of vowel spectra, such as the RF and HMM2 features, are (almost) just as capable as the true formants to discriminate between the vowel classes - even if the features are as crude as HMM2. Especially the results obtained with the HMM2 features, which definitely do not represent formants in the sense of vocal tract resonances, suggest that consistency (including smoothness of the feature tracks over time) is more important than the relation to the underlying, physical speech production process.

The most salient difference between the LDA and HMM results is the classification rates that were obtained for the HMM2 features. While the HMM2 results for the HMM classifier are comparable with the corresponding HLF results, the LDA classifier does not seem to be able to distinguish between the vowel classes if it is trained on HMM2 features. This result indicates that it is not possible to distinguish between the vowel classes in the coarsely sampled HMM2 feature space when only a few points (in time) are taken into consideration. Due to the coarseness of the HMM2 features, HMM2 feature tracks may change rather abruptly at any point in time. For example, an abrupt change may occur before the 20% duration point for some pronunciations of a certain phoneme and after the 20% duration point for other pronunciations of the same phoneme. The LDA classifier does not seem to be able to deal with these differences. The HMM classifier, on the other hand, is able to handle these changes in the data because it classifies vowels in terms of a complete temporal sequence of HMM2 features. However, the coarse quantization of the HMM2 features is not an intrinsic limitation of this approach to the representation of spectral envelopes. On the contrary, it is one of the implications of the way in which the current version of HMM2 has been implemented. Other implementations are presently under investigation, which use filters with much narrower pass bands than the 13 critical band filters used in this study.

In both the LDA and the HMM classification experiments, the classification rates measured for the gender-dependent data sets were higher than the corresponding results for the gender-independent data sets. However, for the HLF data the difference was much smaller than for the other feature representations. It is probably true that the gender-independent HLF data are not truly gender-independent, because the gender of the speakers was known to the human labelers. The HLF features may therefore be said to contain implicit gender information. A comparison of the results obtained with HLF and gender-dependent RF features suggests that the advantage of expert knowledge is rather small when an automatic formant extraction procedure can be configured to avoid errors in assigning spectral peaks to formant numbers.

HLF features could also be expected to have an advantage due to the fact that the labelers knew the phone identities while they were assigning the formant labels. As was pointed out in Section IV, the analysis of formants separately from hypotheses about what is being said will always be prone to errors (Holmes, 2000). The human labelers knew the identity of the tokens they were labeling, i.e. they could use additional information in assigning formant labels. This constitutes another source of implicit knowledge which gives the HLF features an advantage over the automatically derived features: these either rely on imperfect classification results (in the case of HMM2) or have no knowledge about the token for which feature extraction is attempted (in the case of the RF features). Here too the comparison of the HLF and the gender-dependent RF features suggests that the advantage derived from prior knowledge of the vowel identities was not very large. However, this observation may not generalize to other databases, where the phonetic context of the vowels will be richer and have a bigger impact on the spectral envelopes. After all, the /h-V-d/ context was chosen to minimize coarticulation effects, which will be especially cumbersome for automatic (and manual) formant extraction in, for example, the case of nasal consonants.

It is difficult to say to what extent the HLF features relate to formants as resonances of the vocal tract. After all, the experts based their formant measurements on LPC spectra of the radiated speech signals. Although they have used prior knowledge about vocal tract resonances of individual vowels, this knowledge could only be brought to bear on the results in the form of selection of one spectral peak instead of other competing candidates, perhaps even after a change of the LPC order to obtain a peak in the frequency region where it was predicted by acoustic phonetic theory. It is also possible that part of the formant values recorded in the AEV database for vowel onsets and offsets is the result of manual smoothing, interpolation, or extrapolation, again guided by phonetic theory. However, the automatically extracted RF features appeared to resemble the HLF features very closely, provided that the automatic RF extractor was given prior information about the gender of the speaker. Although there is a theoretical relation between LPC spectral estimation and resonances of linear filters that could model the vocal tract (Markel and Gray (Jr.), 1976), and although this relation is enhanced by a proper selection of the LPC order (as was done for the gender-dependent RF extractor), it is now generally accepted that inferences of vocal tract shapes and resonances from spectral envelopes are not well possible. This suggests that HLF and gender-dependent RF features are very similar, in the sense that both represent the spectral envelopes in terms of the locations of the major peaks. The results of this study suggest that, for ASR, it is less important that these peaks should correspond to true vocal tract resonances than that the selection of the peaks in the presence of multiple candidates is done consistently. The fact that consistency is more important than relation to vocal tract resonance is most clearly demonstrated by the power of the HMM2 features, which do not even relate to spectral maxima per se, but which appear to be very consistent.

The results in Section V.E show that the formant-like features that were investigated in this study are not inherently robust against additive noise. Neither of the two representations was able to keep track of the spectral maxima that should remain intact in noisy speech data. The finding that it is not possible to build a successful classifier using features that are inherently error-prone is not very surprising. For the use of formants in ASR the message appears to be that the theoretical advantages of the formant representation are neutralized by the enormous difficulty of building a reliable automatic formant extractor, especially one that is also able to process noisy speech. Until such a powerful formant extractor is available, there seems to be little advantage in adding formant measures to the set of features in ASR. The relative success of adding formant candidates to MFCC parameters in the work of (Holmes et al., 1997) does not contradict this conclusion. After all, their results can be considered as the simultaneous solution of two closely related problems: formant extraction and ASR. For formant extraction there is no doubt that the results should improve as speech recognition improves, since knowledge of the sounds is a powerful knowledge source to guide the classification of spectral peaks as formants. For speech recognition one would expect a similar advantage: an interpretation of the signal in terms of sounds and words that make sense against the background of formant candidates should be more accurate than one that does not.

High performance formant extraction in noisy speech will require a different approach to signal processing than the usual spectral estimators that assume the signal to be stationary over the duration of an analysis window. Several techniques based on models of the signal processing in the mammalian auditory system have been proposed. (Andringa, 2002) is a recent example which is especially interesting because it argues that signal processing and recognition are intimately intertwined. The decision whether a spectral maximum is indeed a vowel formant is made dependent not only on the characteristics of the signal itself (are local spectral peaks consistent with a very precise estimate of the instantaneous fundamental frequency ?) but also on whether a vowel with the hypothesized formant structure could be present at a specific point in the signal. This suggests that, for a formant representation to have its maximum impact on ASR, it is not just the signal processing and feature extraction that must be advanced. Major advances in the search and decision process that eventually link features to words, meanings and intentions are also required.

VII Conclusions

In this paper three issues were investigated within the framework of the AEV database introduced in (Hillenbrand et al., 1995). In the first instance, it was shown that, using standard ASR methods, hand-labeled formants only marginally outperform automatically extracted formant-like features such as RFs and HMM2 features on a vowel classification task.

Secondly, a comparison of hand-labeled formants, RFs and HMM2 features revealed that there is little advantage in using acoustic features that have a direct relation to vocal tract resonance for the classification of vowels. Although gender-dependent RF features resemble hand-labeled formants quite closely, this is not the case for HMM2 features. The latter do not even relate to spectral peaks, but rather to transitions between minima and maxima of the spectral envelope. The most likely explanation for the (small) advantage of hand-labeled formants that emerged from this study is their intrinsic smoothness over time, in conjunction with a very high resilience against consistent mis-alignment between spectral peaks and formant labels.

Thirdly, the theoretical robustness of formant measures against additive noise could not be verified for either of the two automatically extracted formant-like feature sets. The lack of robustness of these features does not necessarily imply the rejection of the hypothesis that the formants remain visible as peaks in the spectral envelope. Rather, the noise seems to introduce additional spectral peaks, which cannot be effectively discarded as formant candidates by the relatively simple signal processing techniques underlying RF extraction and HMM2 feature computation. The theoretical advantages of the formant concept for processing noisy speech can only be harnessed by signal processing techniques that take full profit of continuity and coherence in the signals, both in time and in frequency.

In summary, it is fair to say that in clean conditions the formant representation of speech signals has no compelling advantages over representations that do not involve error-prone labeling decisions (such as MFCCs used in this and many other studies). In noisy conditions the theoretical advantages of the formant concept are vastly diminished by the failure of almost all signal processing techniques to reliably distinguish between spectral maxima that must be attributed to vocal tract resonances and maxima that are introduced by the noise.

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.

Notes

[1]Some of the experimental results reported in this study were presented in “Evaluation of formant-like features for ASR”, Proceedings of the 7^{th} International Conference on Spoken Language Processing, Denver, U.S.A., September 2002.

[2] Vowel steady state was defined by Peterson and Barney as, "... following the influence of the /h/ and preceding the influence of the /d/, during which a practically steady state is reached" (Peterson and Barney, 1952).

[3]The possibility to apply pre-emphasis is incorporated in the acoustic pre-processing of the RF algorithm. One may therefore assume that the inherent spectral tilt in the data is equalized and that all the LPC poles are available to model spectral peaks.

[4]In (Hillenbrand and Gayvert, 1993) it was found that, for a vowel classification task, nonlinear frequency transforms significantly enhanced the performance of a linear discriminant classifier. For a quadratic classifier, on the other hand, there was no advantage for any of the nonlinear transforms (mel, log, Koenig, Bark) over linear frequency. During the current investigation HMM classification experiments were also conducted using the original, linear frequency values. No significant difference was observed between the tests performed with the linear frequency values and the mel-scaled values.

[5]These features were derived using HTK's feature extraction software (Young et al., 1997).

[6]Data from /e/ and /o/ were omitted in (Hillenbrand et al., 1995) to facilitate comparisons with Peterson and Barney's results.

References

Andringa, T. (2002). *Continuity preserving signal processing.* PhD thesis, Rijksuniversiteit Groningen, Groningen, The Netherlands.

Cooke, M., Green, P., Josifovksi, L., and Vizinho, A. (2001). Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, 34:267–285.

de Veth, J., Cranen, B., and Boves, L. (2001). Acoustic backing-off as an implementation of missing feature theory. *Speech Communication*, 34(3):247–265.

de Wet, F., Cranen, B., de Veth, J., and Boves, L. (2000). Comparing acoustic features for robust ASR in fixed and cellular network applications. In *Proceedings of ICASSP 2000*, pages 1415–1418, Istanbul, Turkey.

Flanagan, J. (1972). *Speech Analysis Synthesis and Perception (2nd ed.).* Springer-Verlag, Berlin, Heidelberg, New York.

Garner, P. and Holmes, W. (1998). On the robust incorporation of formant features into hidden Markov models for automatic speech recognition. In *Proceedings of ICASSP 1998*, pages 1–4.

Hillenbrand, J. and Gayvert, R. (1993). Vowel classification based on fundamental frequency and formant frequencies. *Journal of Speech and Hearing Research*, 36:694–700.

Hillenbrand, J., Getty, L., Clark, M., and Wheeler, K. (1995). Acoustic characteristics of American English vowels. *JASA*, 97(5):3099–3111.

Holmes, J., Holmes, W., and Garner, P. (1997). Using formant frequencies in speech recognition. In *Proceedings of Eurospeech 1997*, pages 2083–2086, Rhodes, Greece.

Holmes, W. (2000). Segmental HMMs: Modelling dynamics and underlying structure for automatic speech recognition. http://www.ima.umn.edu/multimedia/fall/m1.html.

Hunt, M. J. (1999). Spectral signal processing for ASR. In *Proceedings of ASRU 1999*, Keystone, Colorado, USA.

Markel, J. and Gray (Jr.), A. (1976). *Linear Prediction of Speech.* Springer-Verlag, Berlin.

Minifie, F., Hixon, T., and Williams, F., editors (1973). *Normal aspects of speech, hearing and language.* Prentice Hall, Englewood Cliffs, New Jersey, USA.

Nadeu, C. (1999). On the filter-bank-based parameterization front-end for robust HMM speech recognition. In *Proceedings of Nokia workshop on robust methods for speech recognition in adverse conditions*, pages 235–238, Tampere, Finland.

Noisex (1990). NOISE-ROM-0. NATO: AC243/(Panel 3)/RSG-10, ESPRIT: Project 2589-SAM.

Peterson, G. and Barney, H. (1952). Control methods used in a study of the vowels. *JASA*, 24:175–184.

Rabiner, L. and Schafer, R. (1978). *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, New Jersey, USA.

Ramakrishnan, B. (2000). *Reconstruction of incomplete spectrograms for speech recognition*. PhD thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.

Renevey, P. and Drygajlo, A. (2000). Statistical estimation of unreliable features for robust speech recognition. In *Proceedings of ICASSP 2000*, pages 1731–1734, Istanbul, Turkey.

Weber, K., Bengio, S., and Bourlard, H. (2000). HMM2 - A novel approach to HMM emission probability estimation. In *Proceedings of ICSLP 2000*, pages (III)147–150, Beijing, China.

Weber, K., Bengio, S., and Bourlard, H. (2001a). HMM2 - extraction of formant structures and their use for robust ASR. In *Proceedings of Eurospeech 2001*, pages 607–610, Aalborg, Denmark.

Weber, K., Bengio, S., and Bourlard, H. (2001b). A pragmatic view of the application of HMM2 for ASR. IDIAP-RR 23, IDIAP, Martigny, Switzerland.

Weber, K., Bengio, S., and Bourlard, H. (2001c). Speech recognition using advanced HMM2 features. In *Proceedings of ASRU 2001*, Madonna di Campiglio, Trento, Italy.

Weber, K., de Wet, F., Cranen, B., Boves, L., Bengio, S., and Boulard, H. (2002). Evaluation of formant-like features for ASR. In *Proceedings of ICSLP 2002*, Denver, U.S.A.

Willems, L. F. (1986). Robust formant analysis. In *IPO Annual report 21*, pages 34–40, Eindhoven, The Netherlands.

Young, S., Odell, J., Ollason, D., Valtchev, V., and Woodland, P. (1997). *The HTK Book (for HTK Version 2.1)*. Cambridge University, Cambridge, UK.

Table I Mean Mahalanobis distance between RF features and hand-labeled data.

gender	RF3	3RF4
male	3.5	2.1
female	1.6	5.3
all	1.9	3.0

Table II LDA classification results: gender-independent data.

Feature type	stst	20%80%	20%stst80%
HLF	77.0 (± 2.5)	91.4 (± 1.7)	91.9 (± 1.6)
RF	63.4	81.8	83.0
HMM2	31.7	48.7	52.2
MFCC13	73.1	90.5	91.2
MFCC3	57.5	78.6	78.2

Table III LDA classification results: gender-dependent data.

Feature type	stst	20%80%	20%stst80%
HLF	79.4 (± 2.4)	93.6 (± 1.5)	93.8 (± 1.4)
RF	76.1	91.2	92.0
HMM2	48.5	60.1	63.8
MFCC13	81.7	94.5	94.2
MFCC3	64.2	84.8	84.9

Table IV HMM classification results for gender-independent and gender-dependent data.

Feature type	Gender-independent	Gender-dependent	Feature dimension
HLF	87.7 ($\pm$2)	89.6 ($\pm$1.8)	6
RF	84.1	90.5	6
HMM2	77.0	87.2	3
MFCC13	92.3	92.1	26
MFCC3	77.6	81.2	6

Figure captions

Figure 1: Left panel: Schematic representation of an HMM2 system in the time/frequency plane. The left-right model is the temporal HMM with a top-down frequency HMM in each of its states. **Right panel:** Example of a 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. The corresponding indices form an HMM2 feature vector (right).

Figure 2: Tracks of HLF, RF and HMM2 features for one female pronunciation of the vowel /ɝ/ projected onto (a) the mel-scaled log-energy of each frame and (b) the corresponding FF features.

Figure 3: Average classification rates (% correct) for gender-dependent models trained on clean MFCC13 (+), RF ($*$) and HMM2 (o) features and tested in babble (left panel) and factory (right panel) noise. The corresponding feature vector dimensions are 26 (MFCC13), 6 (RF) and 3 (HMM2).

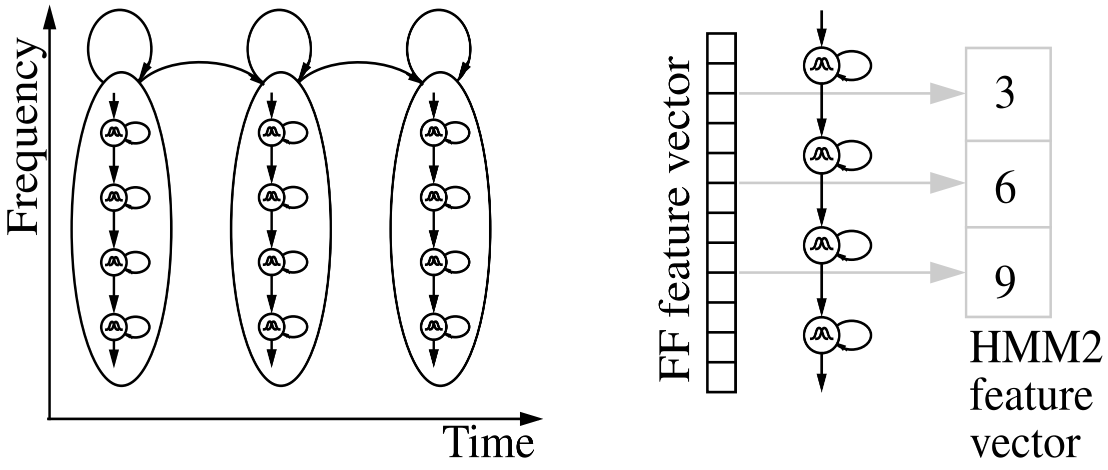
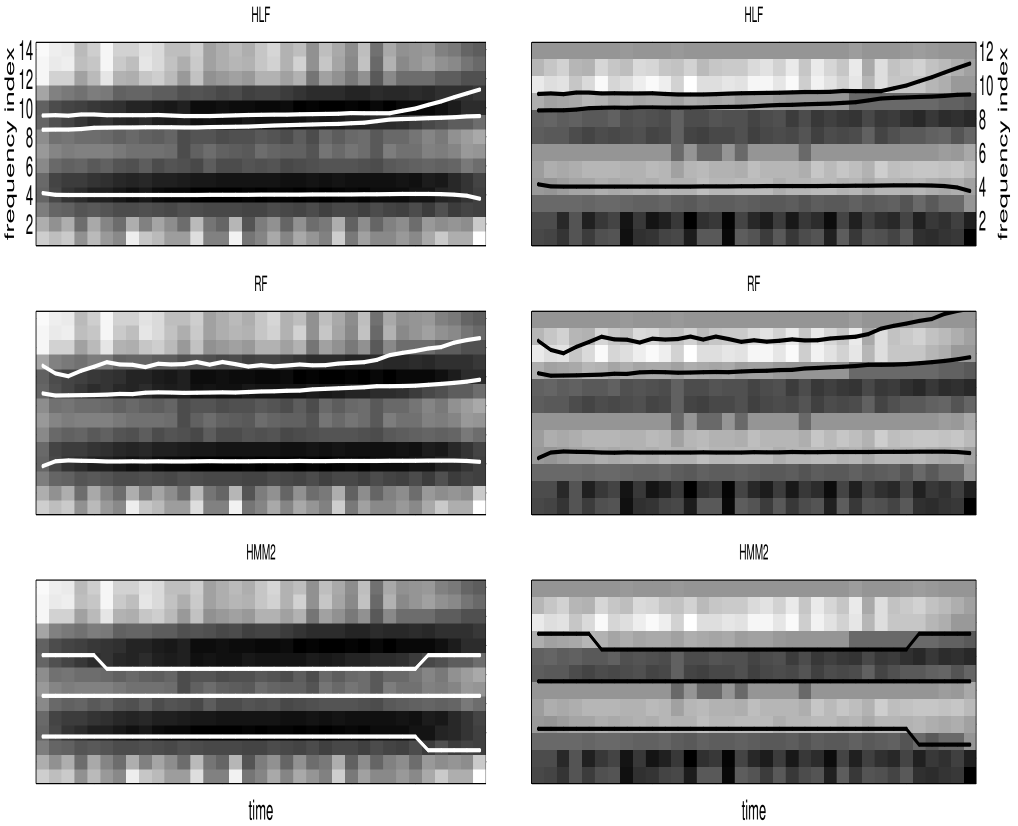


Figure 1


HLF
HLF
frequency index
14
12
10
8
6
4
2
12
10
8
6
4
2
frequency index
RF
RF
HMM2
HMM2
time
time

(a) (b)

Figure 2

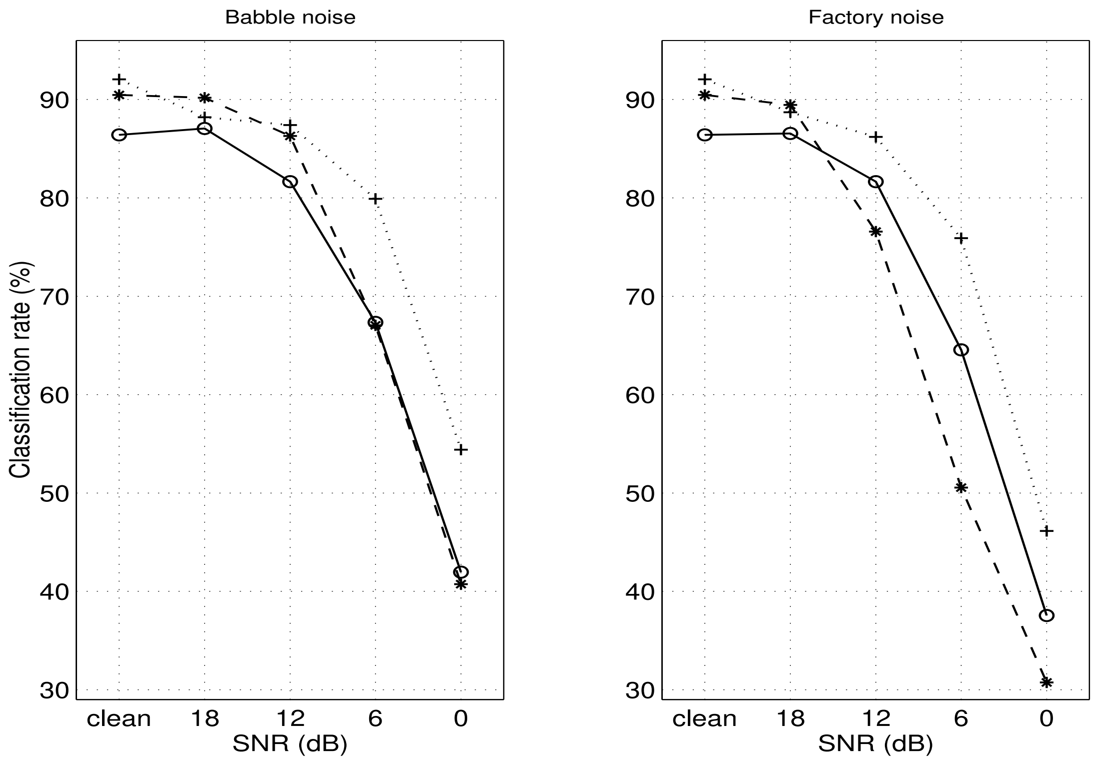

Babble noise
Factory noise
Classification rate (%)
90
80
70
60
50
40
30
clean
18
12
6
0
SNR (dB)

Figure 3