Evaluation of formant-like features on an automatic vowel classification task[1]

Febe de Wet[a)]
Department of Language and Speech, University of Nijmegen, Nijmegen, The Netherlands

Katrin Weber[b)]
IDIAP—Dalle Molle Institute for Perceptual Artificial Intelligence, Martigny, Switzerland and EPFL—Swiss Federal Institute of Technology, Lausanne, Switzerland

Louis Boves[c)] and Bert Cranen[d)]
Department of Language and Speech, University of Nijmegen, Nijmegen, The Netherlands

Samy Bengio[e)]
IDIAP—Dalle Molle Institute for Perceptual Artificial Intelligence, Martigny, Switzerland

Hervé Bourlard[f)]
IDIAP—Dalle Molle Institute for Perceptual Artificial Intelligence, Martigny, Switzerland and EPFL—Swiss Federal Institute of Technology, Lausanne, Switzerland

(Received 20 December 2002; accepted for publication 23 April 2004)

Numerous attempts have been made to find low-dimensional, formant-related representations of speech signals that are suitable for automatic speech recognition. However, it is often not known how these features behave in comparison with true formants. The purpose of this study was to compare two sets of automatically extracted formant-like features, i.e., robust formants and HMM2 features, to hand-labeled formants. The robust formant features were derived by means of the split Levinson algorithm while the HMM2 features correspond to the frequency segmentation of speech signals obtained by two-dimensional hidden Markov models. Mel-frequency cepstral coefficients (MFCCs) were also included in the investigation as an example of state-of-the-art automatic speech recognition features. The feature sets were compared 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 and in noisy acoustic conditions. When using clean data, the classification performance of the formant-like features compared very well to the performance of the hand-labeled formants in a gender-dependent experiment, but was inferior to the hand-labeled formants in a gender-independent experiment. The results that were obtained in noisy acoustic conditions indicated that the formant-like features used in this study are not inherently noise robust. For clean and noisy data as well as for the gender-dependent and gender-independent experiments the MFCCs achieved the same or superior results as the formant features, but at the price of a much higher feature dimensionality. [DOI: 10.1121/1.1781620]

PACS numbers: 43.72.Ne, 43.72.Ar [DOS] Pages: 1781–1792

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. Speech production is often modeled as a source signal feeding into a linear all-pole filter. In terms of this model, the phonetically relevant properties of speech signals are the resonance frequencies of the filter, also known as formants. The formant representation of speech signals is attractive because it is parsimonious yet powerful. For instance, it is well known that the frequencies of the first two or three formants are sufficient for the perceptual identification of vowels (Pols *et al.*, 1969; Flanagan, 1972; Minifie *et al.*, 1973). Many attempts have therefore been made to exploit the formant representation in speech synthesis, speech coding and automatic speech recognition (ASR).

A special reason why formants are attractive 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 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). The formant representation may therefore be expected to be robust against additive noise. Automatically extracted formant-like[2] features have shown some potential

[a)]Electronic mail: F.de.wet@let.kun.nl
[b)]Electronic mail: weber@idiap.ch
[c)]Electronic mail: l.boves@let.kun.nl
[d)]Electronic mail: b.cranen@let.kun.nl
[e)]Electronic mail: bengio@idiap.ch
[f)]Electronic mail: bourlard@idiap.ch

for noise robustness in automatic speech recognition, especially when combined with nonparametric spectral features (Garner and Holmes, 1998; de Wet *et al.*, 2000; Weber *et al.*, 2001a).

Despite its apparent advantages, the formant representation is not widely used in speech technology applications. In this area, nonparametric representations of speech signals are most commonly used. Even if the estimate of the spectral envelope is derived from a parametric estimator such as Linear Predictive Coding (LPC) [which can be related to the source-filter model of acoustic speech production (Markel and Gray, 1976)], speech systems avoid an explicit interpretation of the spectral envelope in terms of formants.

Given the explanatory power of the formant representation in speech production and perception research, its absence in speech technology seems awkward. One of the reasons why formants are not widely used in speech technology is that there is no one-to-one relation between the spectral maxima of an arbitrary speech signal and the resonance frequencies of the vocal tract. The exact causes of the many-to-many mapping between spectral maxima and true formants need not concern us here. What is essential is that despite numerous attempts to build accurate and reliable automatic formant extractors (e.g., Flanagan, 1972; Rabiner and Schafer, 1978; Welling and Ney, 1996; Garner and Holmes, 1998; Bazzi *et al.*, 2003), there are still no tools available that can automatically extract true formants from speech reliably. Labeling spectral maxima as formants is often only possible if the phonetic label of the sound is known, because the spectra may contain a varying number of prominent maxima (Garner and Holmes, 1998; Stevens, 1998).

The many-to-many relation between spectral maxima and true 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 the resonance frequencies of a linear all-pole filter. Nasals and fricatives, for example, can only be accurately described if antiresonances are specified in addition to the resonances (Ladefoged, 1975; Stevens, 1998). The voice source may also contain spectral peaks and valleys that may affect the spectral peaks in the corresponding speech signals. 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 typical speech signals show fewer spectral maxima than the number of vocal tract resonances predicted by acoustic phonetic theory. Most of the search algorithms that are used in ASR are designed to deal with feature vectors of a fixed length. Formant extractors which do not yield a fixed number of spectral peaks labeled as formants for each data frame can therefore not be used in conjunction with standard ASR search algorithms.

If it is difficult, if not impossible, to consistently and reliably extract true formants from arbitrary speech signals, the question arises whether the formant-like parameters that are delivered by one of the existing "formant" extraction techniques are as versatile as the true vocal tract resonances. To be useful for current ASR applications, a formant extractor must be guaranteed to deliver an equal number of formant parameters for each speech frame. Moreover, if the parameter values must have at least some relation to vocal tract resonances, they must develop smoothly over time. In this study two formant-like feature representations that fulfill both these basic requirements were investigated: two-dimensional hidden Markov models (HMM2) (Weber *et al.*, 2000) and robust formants (RFs) (Willems, 1986). The details of these techniques will be explained in Secs. II B and II C.

The best way to compare the performance of automatically extracted formant-like features and true formants would be to evaluate their performance in a real ASR system. However, all state-of-the-art ASR systems rely on very large corpora to train probabilistic models in a fully automatic manner. Obtaining corpora that are sufficiently large for ASR purposes is only feasible if no manual intervention is needed in the acoustic analysis of the signals. Due to the lack of tools to compute true formants reliably and accurately, experts are needed to add formant labels to the speech in a training database. This makes it practically impossible to provide sufficiently large training corpora for the development of formant-based processing. Yet, the theoretical attractiveness of the formant representation has motivated several attempts to overcome this hurdle.

One way to circumvent the problem that there are no databases with true formant labels that are sufficiently large to train an ASR system, is to look for another task on which the representations can be compared, and from which one might draw inferences to realistic ASR tasks. Such a task would, of course, require a suitably labeled database. 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 AEV data have been used for experiments with human and automatic vowel classification, a task that is much simpler than continuous speech recognition. However, it is safe to assume that if a formant-like representation fails to approach the same vowel classification performance as the true formants in the AEV database, it is highly unlikely that such a representation could yield the theoretical advantage expected from true formants on a more realistic continuous speech recognition task.

Thus, the goal of the research reported in this paper was to investigate the degree to which formant-like features can approximate the performance of true formants in a vowel classification task, and to interpret the results in terms of the extent to which formant-like features can harness the theoretical advantages of true formants in ASR. More specifically, the aims of the research reported here are

(1) to investigate the degree to which RFs and HMM2 features resemble true formants.
(2) to compare the performance of true formants with RFs and HMM2 features on a vowel classification task. In order to strengthen the link with current research in ASR, a set of nonparametric features, i.e., mel-frequency cepstral coefficients (MFCCs), was also included in the experiments. In addition, two different classification techniques were used: Linear Discriminant Analysis (LDA) and Hidden Markov Models (HMMs). The outcome of these experiments should indicate to what ex-

tent a close relation between acoustic features and vocal tract resonance frequencies is important for automatic vowel classification.

(3) to investigate the claim that formant-like features are inherently robust against additive noise, because they are related to the spectral maxima that will stay above the local spectral level of additive noise.

The rest of this paper is organized as follows: Section II briefly introduces the AEV database, the RF algorithm, and the HMM2 feature extractor. Section III reports on the experimental setup and the results of the vowel classification experiments. The results are followed by a discussion and conclusions in Secs. IV and V, respectively.

II. DATABASE AND FORMANT EXTRACTION

A. 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) presented in Hillenbrand *et al.* (1995). The AEV database contains recordings of 12 vowels (/i, ɪ, ɛ, æ, ɑ, ɔ, ʊ, u, ʌ, ɝ, e, o/) produced in isolated /h-V-d/ syllables by 45 men, 48 women, and 46 children. Various acoustic measurements were made for each token in the database, including vowel duration, vowel steady-state times,[3] formant tracks, and fundamental frequency tracks.

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 scientists. 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. Only the formant tracks corresponding to the vowel sections of the /h-V-d/ sections were therefore used in the classification experiments described in Sec. III.

Where irresolvable formant mergers occurred, Hillenbrand *et al.* put zeros into the higher of the two formant slots affected by the merger. In order to use the vowels containing mergers for our classification experiments, we replaced the zeros by the frequency value in the lower formant slot, i.e., two equal values were used. Irresolvable mergers occurred in about 4% of the vowel tokens.

In the Hillenbrand study, F1, F2, and F3 were measured for all the signals. 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. For the purpose of the current investigation, we therefore decided to limit the scope of the hand-labeled formant feature set to the first three formants. In addition, we decided to use an equal number of male and female utterances and not to use the children's data. The latter decision was made because it could not be guaranteed that the two automatic formant extractors could handle children's speech appropriately.

The mean values that were measured for the first three male and female formants were all well below 4 kHz (Hillenbrand *et al.*, 1995). We therefore decided to downsample the original 16 kHz speech data to 8 kHz. Furthermore, the acoustic analyses in our experiments adhered to the same time resolution used by Hillenbrand *et al.* 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. Finally, 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] (Davis and Mermelstein, 1980; Rabiner and Juang, 1993).

B. Robust formant algorithm

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 (Delsarte and Genin, 1986). Instead of directly applying a root solving procedure to the LPC polynomial, a so-called singular predictor polynomial is constructed from which the zeros are determined in an iterative procedure. The iterative procedure guarantees that the number of complex conjugate pairs of zeros is always equal to half the LPC order, provided that the order is even. Thus, the algorithm will always return the same number of parameters. Moreover, since the procedure tends to spread the zeros evenly on the unit circle, it enforces a large degree of continuity in the parameter tracks (as a function of time). After the frequency positions of the RF features have been established, their corresponding bandwidths are chosen from a predefined table such that the resulting all-pole filter minimizes the error between the predicted data and the input.

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 the previous section, the scope of this study is limited to the frequency range between 0 and 4 kHz and to the values of the first three formants. However, in the AEV database the mean value (taken over all the relevant data) of F4 is 3.536 kHz (σ=135.5) for males and 4.159 kHz (σ=174.7) for females. This implies that, for some of the vowels produced by male speakers, the frequency band between 0 and 4 kHz may contain four vocal tract resonances instead of three. An automatic formant extraction procedure applied to the AEV data should therefore 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 this requirement is to use a sixth-order LPC analysis.[5] However, the accuracy of the LPC analysis is bound to suffer if a sixth-order analysis is used to analyze spectra with four maxima, because two complex poles are usually required to model each spectral peak (Stevens, 1998). In these cases an eighth-order LPC seems more appropriate, although it introduces the need to select three RFs from the set of four.

TABLE I. Mean Mahalanobis distance between the hand-labeled formants and the RF features.

Gender	RF3	3RF4
Male	3.5	2.1
Female	1.6	5.3
All	1.9	3.0

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 calculated every 8 ms over 16 ms Hamming windowed segments. We subsequently calculated the Mahalanobis distance between the hand-labeled formants (HLFs) and the RF3 and 3RF4 features, respectively. The Mahalanobis distance between two distributions is defined as (Duda *et al.*, 2001):

$$r^2=(\mathbf{x}-\mu)^t\sum{}^{-1}(\mathbf{x}-\mu). \quad (1)$$

The mean Mahalanobis distance (across all vowels) between the HLFs and the two sets of robust formants are given in Table I. The results in Table I show that the RF features are closer to the HLFs if the order of the analysis corresponds to the inherent signal structure. 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 RFs and HLFs increases. 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.

C. The HMM2 feature extractor

HMM2 is a special mixture of hidden Markov models (HMMs), in which the emission probabilities of a conventional, temporal HMM are estimated by a secondary HMM (Weber *et al.*, 2001b). As shown in Fig. 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.

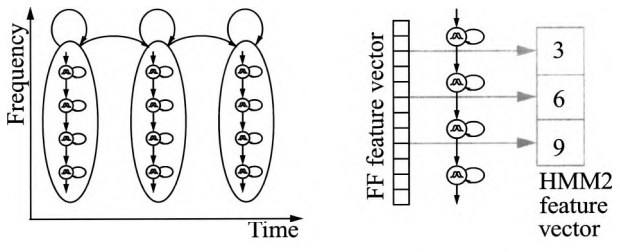

FIG. 1. Left panel: Schematic representation of an HMM2 system in the time/frequency plane. The left–right model is the temporal HMM with a top–down frequency HMM in each of its states. Right panel: Example of a temporal "FF" vector (left) as emitted by a frequency HMM. Each of the squares in this feature vector corresponds to a four-dimensional subvector. Gray 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).

In fact, each temporal feature vector can be seen as a sequence of subvectors. The subvectors 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 subvectors 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 (GMMs). 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.

Frequency filtered filterbanks (FFs) (Nadeu, 1999) are typically used as features for HMM2, because they are comparatively decorrelated in the spectral domain. In certain 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 MFCCs (Nadeu, 1999). For the HMM2 systems that were used in this study, a sequence of 12 FF coefficients was calculated every 8 ms. While a larger number of FF coefficients could possibly be advantageous, this number was chosen in order to make the number of features used for HMM2 comparable to that conventionally used in HMMs. Together with their first- and second-order time derivatives plus an additional frequency index, these FF coefficients form a sequence of 12 four-dimensional subvectors. Each square in the vector labeled "FF feature vector" in Fig. 1 therefore represents a four-dimensional subvector.

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 from the Viterbi segmentation at which point in frequency (i.e., between which subvectors) a transition from one frequency HMM state to the next took place. For example, in Fig. 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 subvector. In the case of four frequency HMM states connected in a top–down topology (as

seen in Fig. 1), we therefore obtain three integer indices (corresponding to precise frequency values). In our classification experiments, these indices were used as three-dimensional feature vectors in a conventional HMM.

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 options 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 indices*, different *initialization* possibilities, as well as different (combinations of) segmentation strategies that can be applied for *training and test* purposes. These design options are discussed in detail in Weber (2003).

The models that were used to obtain the results reported on in Sec. III all had a three-state, left–right topology in the time domain and a four-state top–down topology in the frequency domain. Frequency indices were included as additional feature components in the frequency subvectors. The initialization of the gender-independent HMM2 models was based on the assumption of alternating high and low energy frequency HMM states. 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 on the AEV database (Weber *et al.*, 2002). In that study, it was argued that this result can be attributed to the nature of the AEV data, exhibiting only very few spectral changes (see Sec. III A 2 for a graphical illustration), in conjunction with the very crude nature of the HMM2 features. Often, the frequency segmentation of one phoneme would be the same for all time steps, resulting in zero-valued time derivatives. In other cases, oscillations between two neighboring segmentations were observed, which gave equally meaningless derivatives.

III. EXPERIMENTS AND RESULTS

In the following, the design, execution, and results of the vowel classification experiments are described. In Sec. III A, the first question posed in Sec. I is addressed, i.e., to what extent the features yielded by the two automatic formant extractors resemble the hand-labeled formants in the AEV database. The design of the classification experiments is subsequently described in Sec. III B. Section III C reports on the Linear Discriminant Analysis (LDA) classification results. The LDA experiments enable us to relate our results to those reported in Hillenbrand *et al.* (1995). The results of the HMM classification experiments are presented in Sec. III D.

TABLE II. Mean Mahalanobis distance between the hand-labeled formants, RFs and HMM2 features.

Gender	RF	HMM2
Male	2.1	8.0
Female	1.6	9.1
All	1.9	5.6

The HMM experiments were conducted in order to determine whether the classification performance of hand-labeled formants with LDA generalizes to the classification performance obtained with the maximum likelihood (ML) procedures that are dominant in the ASR community. Finally, Sec. III E reports on the classification performance of the automatically extracted formant-like features in (simulated) noisy acoustic conditions.

A. How formant-like are RFs and HMM2 features?

There are no generally accepted procedures to assess the degree to which formant-like features resemble true formants. In this study we approached the problem in two complementary ways: by means of a formal distance measure that captures the goodness-of-fit in a single measure, and by means of a graphical illustration of the physical nature of the differences that underlie the summary measures.

1. Statistical distance

The Mahalanobis distance was introduced in Sec. II B as a means to select the RF feature sets that were closest to their hand-labeled counterparts, in terms of statistical distance. The minimum values from Table I are repeated in Table II, together with the mean Mahalanobis distances between the HLFs and the HMM2 features. The values in Table II clearly indicate that, in terms of statistical distance, the RFs are more similar to the HLFs than the HMM2 features.

2. Graphical illustration

Some of the issues involved in comparing HLFs, RFs, and HMM2 features can be illustrated by means of a typical example, in the form of a representative token of the vowel /ɜ/. Figure 2 shows the HLF tracks corresponding to a female pronunciation of the vowel overlaid on a spectrogram representing the frequency range between 0 and 4000 Hz.

The same example was used to create the graphs in Fig. 3. In each of the subplots in Fig. 3, the *y* axis corresponds to frequency index, the *x* axis to time, and darker shades of gray to higher intensity levels. Figure 3(a) shows the same HLFs as in Fig. 2, but overlaid on the mel-weighted log-energy within each frame. The mel-scaled filterbank that was used to obtain the energy values consisted of 14 filters that were linearly spaced in the mel frequency domain between 0 and

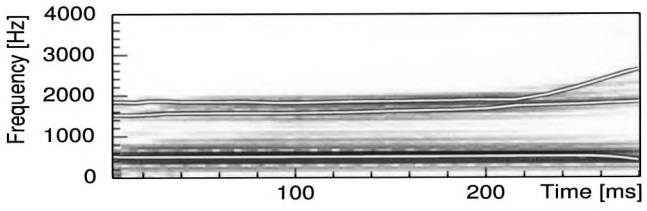

FIG. 2. HLF tracks corresponding to a female pronunciation of the vowel /ɜ/.

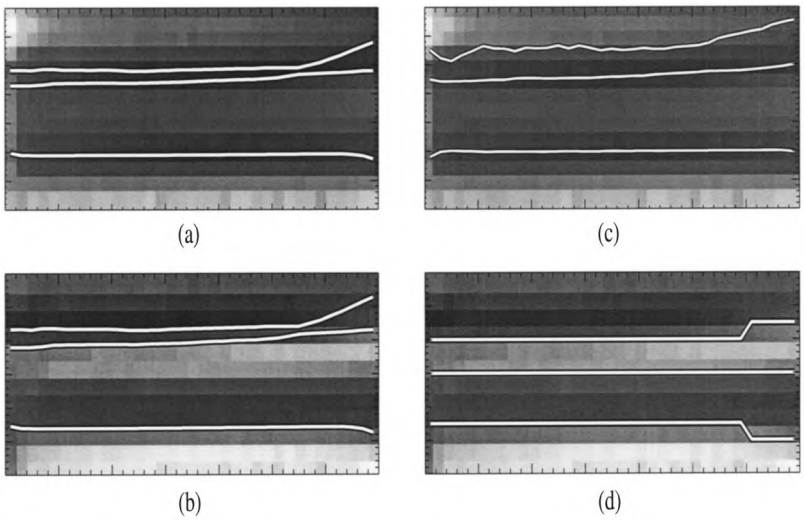

FIG. 3. Feature tracks corresponding to a female pronunciation of the vowel /ɜ/: (a) HLFs overlaid on the mel-scaled log-energy of each frame, (b) the same HLFs on the corresponding FF features; (c) RFs on the mel-scaled log-energy of each frame, and (d) HMM2 feature tracks on FF features.

2146 mel (corresponding to the frequency range between 0 and 4000 Hz, as in Fig. 2). Figure 3(b) shows the HLFs overlaid on 12 FF features, which were derived form the 14 filterbank values, and which were used to train the HMM2 feature extractor. It can be seen that, while the HLFs follow spectral maxima in the filterbank domain, they are positioned at the transitions from low to high intensity regions in the FF domain. Figure 3(c) shows the tracks of the RF features overlaid on the mel-scaled filterbank features, while Fig. 3(d) shows the HMM2 feature tracks overlaid on FF features.

The data in Fig. 3 show that the RF feature tracks are fairly similar to the HLFs. Most importantly, there are no obvious examples of missing formants or wrong labels. The RF features exhibit more frame-to-frame variation than their hand-labeled counterparts. In this example, the LPC spectrum of the vowel contained multiple peaks in the F2–F3 region, while the human labelers consistently preferred a peak at a lower frequency than the RF procedure. We have not been able to verify whether this type of frame-to-frame variation is related to those parts of the vowels in which the human labelers found it most difficult to find the "correct" spectral peaks. It is also not clear whether this variation has affected the classification performance of the RF features, relative to the more smooth HLF features. During normal human speech the articulators move relatively slowly. The smooth HLF feature tracks therefore seem to be more plausible than the slightly more "noisy" RF features. The short-term variations in the RF features are the result of the attempt of the low-order LPC analysis to account for the spectral envelope in the original acoustic signal, which is not only determined by the vocal tract resonances, but also by the excitation. For the RF extractor to yield feature tracks as smooth as the HLF an additional smoother would have to be applied to the raw RF values.

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. However, the feature tracks in Fig. 3(d) indicate that, for the example utterance illustrated in the figure, the HMM2 method succeeded in separating high from low intensity regions in the FF domain. While the first and third HMM2 feature tracks are roughly situated near formant positions (corresponding to the transition between low and high intensity in the FF domain, and to spectral peaks in the spectrogram), the HMM2 track in the middle can be supposed to correspond to a spectral valley. 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.

For other data examples, unexpected transitions and oscillations of the HMM2 feature tracks were also observed. These effects are explained in more detail in Weber (2003). However, for most examples, one or two HMM2 feature tracks correspond to a certain degree (given their low accuracy, which is limited by the low frequency resolution of the FF features) to the HLFs, while another one frequently corresponds to a spectral valley.

B. Experimental setup

Given the fact that the AEV database is quite small, a three-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, gender-independent and gender-dependent. The gender-independent data sets were defined as three nonoverlapping 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 separately. Each train/test set consisted of 30(train)/15(test) speakers. For the gender-independent data sets, the classification results reported in the following 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 were used to conduct the vowel classification experiments, i.e., hand-labeled formants

TABLE III. LDA classification results (% correct): gender-independent data.

Feature type	stst	20%80%	20%stst80%
Hillenbrand	81.0	91.6	91.8
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
MFCC12	73.1	90.5	91.2
MFCC-LDA3	67.3	88.3	90.1

(HLFs), robust formants (RFs), HMM2 features, and two sets of mel-frequency cepstral coefficients (MFCCs). The MFCCs were included as an example of acoustic features that are commonly used in ASR applications. MFCCs describe the spectral envelope in a small number of orthogonal coefficients (Davis and Mermelstein, 1980; Rabiner and Juang, 1993). Usually, 10 to 15 MFCCs are needed to obtain a sufficiently accurate description of the spectrum. The first set of MFCCs that was used in this study, MFCC12, consisted of 12 MFCCs ($c_1 \ldots c_{12}$).[6] However, the MFCC12 feature set contains four times as many coefficients as the HLF, RF, and HMM2 representations. We therefore decided to create a three-dimensional MFCC set, MFCC-LDA3, by projecting the twelve-dimensional MFCCs into a three-dimensional feature space. In order to accomplish the transformation, an appropriate transformation matrix was derived from the relevant training data by means of LDA.

C. LDA classification results

This section reports on an experiment that compares the performance of RFs, HMM2, and MFCC features to the performance of HLF features on a task that is very similar to the one described in Hillenbrand *et al.* (1995). In contrast with the original study, we used an LDA [instead of quadratic discriminant analysis (QDA)], we included all vowels,[7] and we used only the adult speakers' data. To maintain the equivalence between the LDA experiments described here and the corresponding experiments with HMMs that are described in Sec. III D, we used the three-fold cross-validation scheme described in Sec. III B for training and testing (instead of a leave-1-out jackknifing procedure). As in Hillenbrand's study, we investigated classification performance for a single set of formant values determined in the vowel steady state (stst), pairs of formant values measured at 20% and 80% of the vowel duration (20%80%), and triplets in which the steady state value was added to the values at 20% and 80% of the vowel duration (20%stst80%).

The classification rates obtained for the gender-independent data are given in Table III and those for the gender-dependent data in Table IV. Table III also contains the results from the QDA experiments reported in Hillenbrand *et al.* (1995). The results show that our results for the HLF features, obtained with a simpler discriminant analysis technique, are very close to Hillenbrand's results. Human classification for the same data (based on the complete /h-V-d/ utterances) was 95.4% correct (Hillenbrand *et al.*, 1995). The results in Tables III and IV 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.

TABLE IV. LDA classification results (% correct): 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
MFCC12	81.7	94.5	94.2
MFCC-LDA3	73.9	92.3	93.5

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 parentheses. The values in Tables III and IV show that, with the exception of the MFCC12 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 difference is especially evident for the RF features: for the gender-independent experiments the HLF features outperform the RF features by more than 10% (absolute), whereas the corresponding difference for the gender-dependent experiments is less than 3% (absolute).

The data in Tables III and IV also show that the classification performance of the HMM2 features is substantially lower than the results obtained for the other feature sets. This observation indicates that 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 MFCC12 features achieve classification rates that 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, three HLF features and 12 MFCCs are equally able to discriminate between the vowel classes. The three-dimensional MFCCs outperform both the RFs and the HMM2 features and their classification performance is only slightly inferior to the classification rate achieved by the MFCC12 features.

D. HMM classification rates on clean data

The classification rates in Tables III and IV were obtained by means of a 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 ML optimization criterion. The training algorithms therefore learn the distribution of the data without paying particular attention to the boundaries

TABLE V. HMM classification results (% correct) for gender-independent and gender-dependent data.

Feature type	Gender-independent	Gender-dependent	Feature dimension
HLF	87.7 (±2)	89.6 (±1.8)	6
RF	84.1	90.5	6
HMM2	77.0	87.2	3
MFCC13	92.3	92.1	26
MFCC-LDA3	79.9	81.6	6

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 (e.g., Juang *et al.*, 1996). The LDA classification described in Sec. III C 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.

Toward this aim, we compared the vowel classification rates achieved by the five feature sets used in the LDA experiments. With the exception of the HMM2 features, the first-order time derivatives of all the features were also included in the acoustic feature vectors. Since in mainstream ASR it is usual to add overall energy to MFCC features, we extended the MFCC12 vectors to MFCC13 by adding c_0. The resulting feature vector dimensions for the HLF, RF, HMM2, MFCC13, and MFCC-LDA3 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 Sec. III B. 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 ten continuous density Gaussian distributions. The results of the classification experiments are shown in Table V. Once again, the 95% confidence intervals corresponding to the HLF results are indicated in parentheses. The values in the last column of Table V correspond to the dimensions of the different feature sets.

According to the results in Table V, 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.

A remarkable difference between the LDA and HMM experiments is the excellent classification rate achieved by the HMM2 features: these features perform much better in combination with HMMs than with LDA. Table V shows that, for the gender-dependent data, the HMM2 features not only outperform the MFCC-LDA3s but also approximate the performance of the HLF and RF features, in spite of their lower feature dimensionality.

The data in Table V 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 dimensionality. MFCCs with the same dimension (MFCC-LDA3) perform significantly worse than both MFCC13 and HLF. In contrast to what was observed for the LDA experiments, the RFs and HMM2 features also perform much better in comparison with the MFCC-LDA3 features.

A comparison between the gender-independent and gender-dependent results shows 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. Although not to the same extent as the formant-like features, the performance of the MFCC-LDA3 features is also enhanced by using gender-specific modeling. Only the performance of the MFCC13 features seems to be insensitive to gender differences. The MFCC13 features are probably less sensitive to the gender-dependent properties of the data because, in addition to information on the formants, they also contain information about spectral level and general spectral shape.

E. HMM classification rates on noisy data

In this experiment, the models trained on the MFCC13, RF, and HMM2 features that were used for the experiments described in Sec. III D, were tested in noise. The HLF features could not be included in this experiment, because it was not possible to obtain hand-labeled formants for the noisy data. The models were trained on clean data only and 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). The Noisex babble noise contains speech from many different people speaking simultaneously and individual speakers and utterances cannot be discerned from the hubbub. As a result, the signal power is fairly constant and the long-term spectrum is quite flat. The long-term spectrum of the Noisex factory noise also does not exhibit any significant peaks. However, the factory noise is not stationary; it contains a number of hammer blows and other noise bursts.

Figure 4 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.)

In Sec. 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 comparatively robust against additive noise. However, the

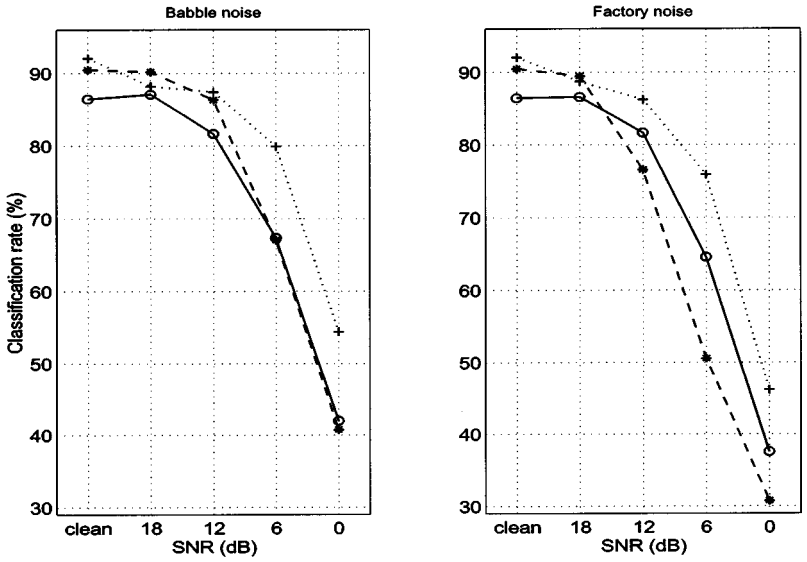

FIG. 4. Average classification rates (% correct) for gender-dependent models trained on clean MFCC13 (+), RF (*), and HMM2 (○) features and tested in babble (left panel) and factory (right panel) noise. The corresponding feature vector dimensions are 26 (MFCC13), 6 (RF), and 3 (HMM2).

results in Fig. 4 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, at lower SNRs the performance degradation of the MFCC13 features seems less severe than that of the formant-related features. To a certain extent, this result may be explained by the fact that the MFCC13 system has a total of 26 feature components at its disposal, while the dimensionality of the RF and HMM2 systems is restricted to 6 and 3, respectively. The higher-order MFCCs—which may contain redundant information in clean conditions—seem to be better at maintaining system performance in adverse acoustic conditions. However, no analysis was done to determine to what extent the errors made when using the different features are complementary. It was also not investigated whether classification performance could be improved by using a combination of different feature streams.

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 Fig. 4 is equivalent to results obtained for other databases in comparable simulations of noisy 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 of an all-pole signal) clearly fail in noisy acoustic conditions. This observation suggests that the RF algorithm is "misled" by the added noise, such that it is no longer capable to find the spectral maxima that correspond to the formants. The noisier the signal, the more the all-pole character of the speech signal disappears. Consequently, the fixed order all-pole model of the RF-algorithm is no longer able to estimate the parameters of the underlying speech production system, and the RF-extractor is turned into a parametric estimator of the peaks in a spectral envelope, the details of which are increasingly determined by the noise.

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.

IV. DISCUSSION

One of the aims of this study was to investigate the degree to which RFs and HMM2 features resemble true formants. The statistical distances and graphical illustrations provided in Sec. III A showed that, of the two automatically extracted formant-like feature sets, the RFs are more similar to the HLFs than the HMM2 features. In fact, the automatically extracted RF features resembled the HLF features quite closely, provided that the RF algorithm was given prior information about the gender of the speaker. This information helps the RF algorithm to avoid one of the most important errors in automatic formant assignment, i.e., labeling spurious peaks as formants, with the results that all higher-order formants in the frame are labeled incorrectly.

Although HMM2 can, in principle, be used as an estimator of true formants, the implementation of HMM2 that was used in this study is not a formant extractor in the classical sense. Because the HMM2 features were derived from a 12-parameter frequency filtered filterbank, they are inherently very coarse. However, the coarse quantization of the HMM2 features is not an intrinsic limitation of this approach to the representation of spectral envelopes. Rather, it is one of the implications of the way in which the current version of HMM2 has been implemented. Other implementations, which use filters with much narrower pass bands than the 14 critical band filters used in this study, should be investigated.

In our comparison of the performance of true formants, RFs, and HMM2 features on a vowel classification task, the

following observations were made. For the gender-dependent data the overall classification performance obtained for the 20%stst80% condition with LDA is better than the results of the HMM classifiers. For the gender-independent data the difference is not equally clear. Apparently, removing the overlap between different vowels from males and females helps the LDA to find an optimal class separation. The ML classifier implemented by the HMMs seems to be less powerful in this regard.

The most salient difference between the LDA and HMM results concerns 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 quantized 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.

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. Classification performance is determined by two factors, i.e., the degree of estimation noise in the features and the overlap between the vowels in the feature space. The observation that the automatically extracted formant-like features generally yielded much better results for the gender-dependent data sets 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 suggests that, for the RF and HMM2 features, the gender separation also achieved a certain degree of reduction in estimation noise in the features themselves.

The classification experiments also showed that the difference between the gender-dependent and gender-independent results was much smaller for the HLF features than for the other feature representations. This observation can probably be explained by the fact that the human labelers knew the gender of the speakers. The labelers also knew the identity of the tokens while they were assigning the formant labels. This gives the HLF features another 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). However, a comparison of the results obtained with HLF and gender-dependent RF features suggests that, for the vowel classification task investigated in this study, the advantage of expert knowledge is rather small when the gender of the speakers is taken into account by the automatic feature extraction procedures. This observation may not generalize to other databases. Especially in fluent, continuous speech 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, e.g., in the case of nasal consonants.

A comparison of the classification performance of HLFs and RFs for the LDA and HMM experiments, and of HLFs and HMM2 features for the HMM experiments, suggests that features that are directly related to vocal tract resonances have very few advantages over formant-like features, as long as the measurement errors in the different feature types are comparable. 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. This result suggests that the formant extraction technique that was recently proposed in Bazzi *et al.* (2003), which guarantees a fixed number of formant values for each frame as well as smooth feature tracks over time, would be a viable candidate to deliver formant-like features that can be used in ASR.

Finally, the results in Sec. III E show that the formant-like features that were investigated in this study are not inherently robust against additive noise. Neither the RFs nor the HMM2 features were able to keep track of the spectral maxima that should remain intact in noisy speech data. 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. 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.

The relative success of adding formant candidates to MFCC parameters in the work of Holmes *et al.* (1997) suggests that a feasible alternative would be to address formant extraction and ASR simultaneously. Hypotheses about formant values should be conditioned by phone observation probabilities, because knowledge of the recognized sound is a powerful knowledge source to guide the classification of spectral peaks as formants. At the same time, an interpretation of the signal in terms of sounds and words that makes sense against the background of formant candidates should result in more accurate ASR than one that does not. 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 processes that eventually link features to words, meanings, and intentions are also required.

V. CONCLUSIONS

In this paper, a number of issues related to the use and usefulness of the formant concept in ASR were investigated. Because there are no databases available that contain enough true formant data to train ASR systems, we focused on the AEV database introduced in Hillenbrand *et al.* (1995).

The first conclusion that can be drawn from our data is that, of the two automatic formant extraction techniques under investigation, robust formants did approximate hand-labeled formants rather closely, provided that the RF algorithm had prior knowledge of the speaker gender. The HMM2 features, on the other hand, did not resemble vocal tract resonances.

Second, for the automatic classification of vowels, we found little advantage in using acoustic features that have a direct relation to vocal tract resonances. If the features are consistent and feature tracks are smooth, their performance can approximate that of true, hand-labeled formants.

Third, the theoretical robustness of formant measures against additive noise could not be verified for either of the two automatically extracted, formant-like feature sets. Background noise seems to introduce additional spectral peaks in the spectral envelopes, which cannot be effectively discarded as formant candidates by the relatively simple signal processing techniques underlying RF extraction and HMM2 feature computation.

In summary, it seems fair to say that, for the clean experimental conditions that were studied in this investigation, the formant representation of speech signals has no compelling advantages (when used as a conventional feature set) over representations that do not involve error-prone labeling decisions such as MFCCs. In noisy conditions, we found that the theoretical advantages of the formant concept were vastly diminished by the failure of our 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.

ACKNOWLEDGMENTS

We would like to thank Professor 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). F.d.W.'s visit to IDIAP was made possible by the I.B.M. Frye grant. K.W. was funded through the Swiss National Science Foundation, Project No. FN 2000-059169.99/1.

[1]Some of the experimental results reported in this study were presented in "Evaluation of formant-like features for ASR," Proceedings of the Seventh International Conference on Spoken Language Processing, Denver, CO, September 2002.

[2]In this paper the term *formant* or *true formants* refers to the resonance frequencies of the vocal tract. The term *formant-like* refers to features that are similar, but not necessarily identical, to true formants.

[3]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).

[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]The possibility to apply pre-emphasis is incorporated in the acoustic preprocessing 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.

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

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

Bazzi, I., Acero, A., and Deng, L. (**2003**). "An expectation maximization approach for formant tracking using a parameter-free non-linear predictor," in Proceedings of ICASSP 2003, Hong Kong, pp. I.464–I.467.

Davis, S. B., and Mermelstein, P. (**1980**). "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Acoust., Speech, Signal Process. **ASSP-28**, 357–366.

Delsarte, P., and Genin, Y. V. (**1986**). "The Split Levinson Algorithm," IEEE Trans. Acoust., Speech, Signal Process. **ASSP-34**, 470–478.

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, Istanbul, Turkey, pp. 1415–1418.

Duda, R. O., Hart, P. E., and Stork, D. G. (**2001**). *Pattern Classification*, 2nd ed. (Wiley, New York).

Flanagan, J. L. (**1972**). *Speech Analysis Synthesis and Perception*, 2nd ed. (Springer, Berlin).

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, Seattle, WA, pp. 1–4.

Hillenbrand, J. M., and Gayvert, R. T. (**1993**). "Vowel classification based on fundamental frequency and formant frequencies," J. Speech Hear. Res. **36**, 694–700.

Hillenbrand, J. M., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**, 3099–3111.

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

Hunt, M. J. (**1999**). "Spectral signal processing for ASR," in Proceedings of ASRU 1999, Keystone, CO.

Juang, B.-H., Chou, W., and Lee, C.-H. (**1996**). "Statistical and discriminative methods for speech recognition," in *Automatic Speech and Speaker Recognition, Advanced Topics*, edited by C.-H. Lee, F. Soong, and K. Paliwal (Kluwer Academic, Boston).

Ladefoged, P. (**1975**). *A Course in Phonetics* (Harcourt Brace Jovanovich, New York).

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

Minifie, F. D., Hixon, T. J., and Williams, F., editors (**1973**). *Normal Aspects of Speech, Hearing and Language* (Prentice–Hall, Englewood Cliffs, NJ).

Nadeu, C. (**1999**). "On the filter-bank-based parametrization front-end for robust HMM speech recognition," in Proceedings of Nokia Workshop on Robust Methods for Speech Recognition in Adverse Conditions, Tampere, Finland, pp. 235–238.

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

Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184.

Pols, L. C. W., van der Kamp, L. J. T., and Plomp, R. (**1969**). "Perceptual and physical space of vowel sounds," J. Acoust. Soc. Am. **46**, 458–467.

Rabiner, L. R., and Juang, B. H. (**1993**). *Fundamentals of Speech Recognition* (Prentice–Hall, Englewood Cliffs, NJ).

Rabiner, L. R., and Schafer, R. W. (**1978**). *Digital Processing of Speech Signals* (Prentice–Hall, Englewood Cliffs, NJ).

Stevens, K. N. (**1998**). *Acoustic Phonetics* (MIT, Cambridge, MA).

Weber, K. (**2003**). "HMM mixtures (HMM2) for robust speech recognition," PhD thesis, Swiss Federal Institute of Technology Lausanne (EPFL), Lausanne, Switzerland.
Weber, K., Bengio, S., and Bourlard, H. (**2000**). "HMM2—A novel approach to HMM emission probability estimation," in Proceedings of ICSLP 2000, Beijing, China, pp. (III)147–150.
Weber, K., Bengio, S., and Bourlard, H. (**2001a**). "HMM2—extraction of formant structures and their use for robust ASR," in Proceedings of Eurospeech 2001, Aalborg, Denmark, pp. 607–610.
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 Bourlard, H. (**2002**). "Evaluation of formant-like features for ASR," in Proceedings of ICSLP 2002, Denver, CO.
Welling, L., and Ney, H. (**1996**). "A model for efficient formant estimation," in Proceedings of ICASSP 1996, Atlanta, GA, pp. 797–800.
Willems, L. F. (**1986**). "Robust formant analysis," in IPO Annual Report 21, Eindhoven, The Netherlands, pp. 34–40.
Young, S., Odell, J., Ollason, D., Valtchev, V., and Woodland, P. (**1997**). *The HTK Book (for HTK Version 2.1)* (Cambridge University, Cambridge).