Formant Estimation for Speech Recognition

Lutz Welling and Hermann Ney, *Member, IEEE*

***Abstract*— This paper presents a new method for estimating formant frequencies. The formant model is based on a digital resonator. Each resonator represents a segment of the short-time power spectrum. The complete spectrum is modeled by a set of digital resonators connected in parallel. An algorithm based on dynamic programming produces both the model parameters and the segment boundaries that optimally match the spectrum.**

We used this method in experimental tests that were carried out on the TI digit string data base. The main results of the experimental tests are: 1) the presented approach produces reliable estimates of formant frequencies across a wide range of sounds and speakers; and 2) the estimated formant frequencies were used in a number of variants for recognition. The best set-up resulted in a string error rate of 4.2% on the adult corpus of the TI digit string data base.

***Index Terms*— Formants, linear prediction, speech analysis, speech recognition.**

I. Introduction

AN EFFICIENT and compact representation of the time-varying characteristics of speech offers potential benefits for speech recognition. Therefore, a variety of approaches such as formant tracking [8], [17], [25], [27], articulatory models [24], and auditory models [13] have been explored. For formant tracking, methods based on linear prediction analysis (LPC) have received considerable attention [20], [26]. Root-finding algorithms are employed to find the zeros of the LPC polynomial, or local maxima of the LPC envelope are searched using peak-picking techniques. The problem with root-finding algorithms is that the determination of formant frequencies and bandwidths is only successful for complex-conjugate poles and not for real poles. Peak-picking techniques are vulnerable to merged formants and spurious peaks.

The approach described in this paper avoids the above-mentioned problems. In [1] and [16], a set of digital formant resonators connected in parallel or in cascade has been proposed for speech synthesis. In this paper, we propose to use a parallel digital resonator model for formant estimation. We model the power spectrum by K formant models, each of which represents one segment of the power spectrum [28]. For the formant model, we use the resonance frequency that is different from the pole frequency typically used in the context of digital resonators. An algorithm based on dynamic programming produces the set of formant parameters and segment boundaries that optimally match the short-time power spectrum of a speech segment. This algorithm is simple to implement and fast. A systematic evaluation of the method on the complete adult corpus of the TI digit string data base [18] is carried out. Formants have also been estimated on the same database in [5] and [17]. We use the estimated formant contours to perform systematic recognition experiments on the TI digit string data base.

An advantage of the method for formant estimation presented in this paper is that an explicit smoothing of the formant frequencies along the time axis does not seem to be necessary, since the formant contours as obtained by the proposed method are remarkably smooth. Other systems such as [27] use frequency continuity constraints and dynamic programming along the time axis in order to get smooth trajectories.

A similar, dynamic programming-based algorithm was presented in [6]. The authors used the algorithm in the context of spectral estimation in order to minimize the discrepancy between a signal spectrum and a model spectrum. Therefore, apart from the segmentation algorithm itself, there are no similarities to the work presented here.

Today, virtually all high-performance speech recognition systems are based on some kind of mel-cepstral coefficients or filterbank analysis. So, we do not expect that in the near future formant-based parameters will be competitive. Nevertheless, there might be specific aspects due to which formant-based parameters are attractive, as listed below.

- Formants are considered to be robust against channel distortions and noise.
- Formant parameters might provide a means to tackle the problem of a mismatch between training and testing conditions.
- There is a close relation of formant parameters to model-based approaches to speech perception and production.

The paper is organized as follows. Section II defines the formant model. Section III describes the dynamic programming algorithm that produces the optimum set of segment boundaries. Section IV contains various experimental results including recognition tests. Finally, our findings are summarized in Section V.

II. Definition of the Formant Model

In this section, we present a model for formant estimation that is based on a set of parallel digital resonators. The frequency range is divided into a fixed number of segments, each of which represents a formant. For the moment, the segment boundaries are fixed. In Section III, we will show how the segment boundaries can be obtained by dynamic programming optimization.

Manuscript received April 17, 1996; revised March 11, 1997. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Mazin Rahim.

The authors are with the Department of Computer Science, Aachen University of Technology, D-52056 Aachen, Germany (e-mail: welling@informatik.rwth.aachen.de; ney@informatik.rwth.aachen.de).

Publisher Item Identifier S 1063-6676(98)00097-2.

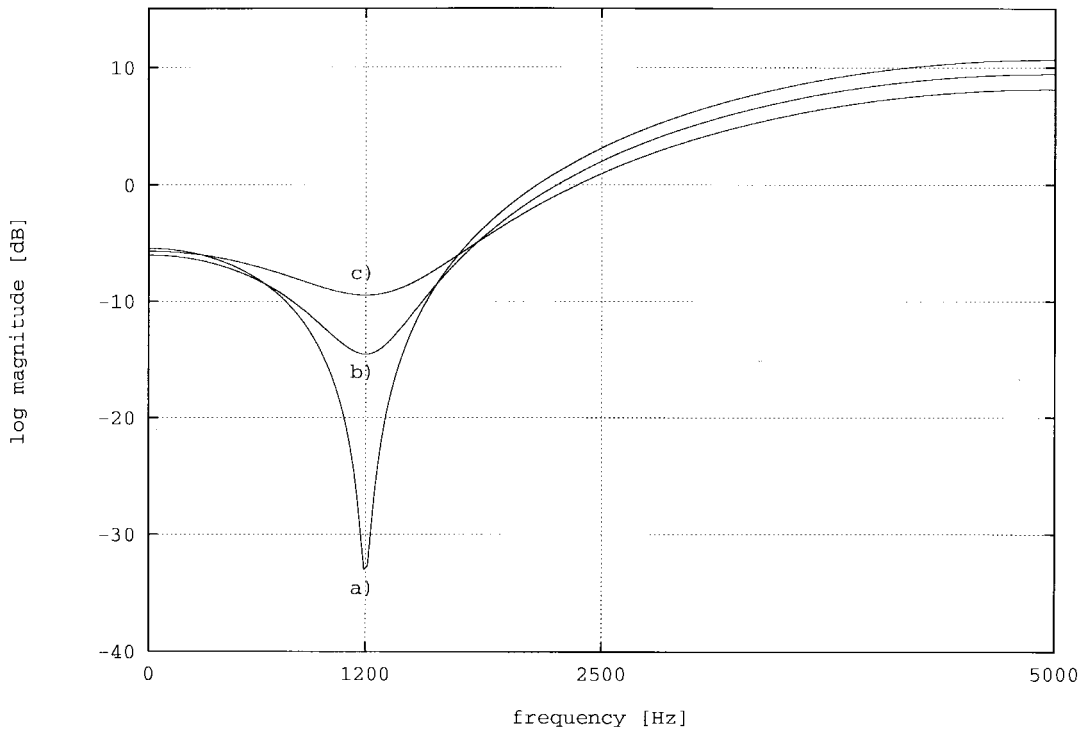


Fig. 1. Predictor polynomial as a function of frequency for a bandwidth of (a) 50 Hz, (b) 500 Hz, and (c) 1000 Hz and a resonance frequency of 1200 Hz.

A. Second-Order Resonator

For each segment k with given boundaries, we define a second-order digital resonator. As in general LPC analysis [23, p. 399], we consider the predictor polynomial, which is defined as the Fourier transform of the corresponding second-order predictor

$$A_k(e^{j\omega}) = 1 - \alpha_k e^{j\omega} - \beta_k e^{j2\omega}$$

where α_k and β_k are the real-valued prediction coefficients. $|A_k(e^{j\omega})|^2$ can be written as

$$\begin{aligned} |A_k(e^{j\omega})|^2 &= 1 + \alpha_k^2 + \beta_k^2 - 2\alpha_k(1-\beta_k)\cos\omega \\ &\quad - 2\beta_k\cos(2\omega) \qquad (1) \\ &= (1+\beta_k)^2 + \alpha_k^2 + \frac{\alpha_k^2(1-\beta_k)^2}{4\beta_k} \\ &\quad - 4\beta_k\left[\cos\omega + \frac{\alpha_k(1-\beta_k)}{4\beta_k}\right]^2. \qquad (2) \end{aligned}$$

The typical frequency dependence of such a predictor polynomial is depicted in Fig. 1. As we will see later, we have the constraint $\beta_k < 0$. Equation (2) shows that the parameter β_k determines the bandwidth of the resonator which is defined [1, p. 128] as the negative logarithm of $(-\beta_k)$. $|A_k(e^{j\omega})|^2$ has its global minimum at the resonance or formant frequency φ_k

$$\varphi_k = \arccos\left[-\frac{\alpha_k(1-\beta_k)}{4\beta_k}\right]. \qquad (3)$$

We denote the beginning point and the end point of segment k by ω_{k-1} and ω_k, respectively. Using the predictor polynomial, we define the prediction error as follows:

$$E(\omega_{k-1}, \omega_k|\alpha_k, \beta_k) = \frac{1}{\pi}\int_{\omega_{k-1}}^{\omega_k} |S(e^{j\omega})|^2\,|A_k(e^{j\omega})|^2\,d\omega$$

where $|S(e^{j\omega})|^2$ denotes the short-time power density spectrum of the speech signal. Using (1), the prediction error can be rewritten as

$$\begin{aligned} E(\omega_{k-1}, \omega_k|\alpha_k, \beta_k) &= (1+\alpha_k^2+\beta_k^2)r_k(0) \\ &\quad - 2\alpha_k(1-\beta_k)r_k(1) - 2\beta_k r_k(2) \qquad (4) \end{aligned}$$

with the autocorrelation coefficients $r_k(\nu)$ of segment k for $\nu = 0, 1, 2$

$$\begin{aligned} r_k(\nu) &:= r_{(\omega_{k-1},\omega_k)}(\nu) \\ &= \frac{1}{\pi}\int_{\omega_{k-1}}^{\omega_k} |S(e^{j\omega})|^2\cos(\nu\omega)\,d\omega. \qquad (5) \end{aligned}$$

By minimizing the prediction error as given by (4) with respect to α_k and β_k, we obtain the following optimum prediction coefficients [19, p. 568]:

$$\begin{aligned} \alpha_k^{\text{opt}} &= \frac{r_k(0)r_k(1) - r_k(1)r_k(2)}{r_k(0)^2 - r_k(1)^2} \\ \beta_k^{\text{opt}} &= \frac{r_k(0)r_k(2) - r_k(1)^2}{r_k(0)^2 - r_k(1)^2}. \end{aligned}$$

The value of the minimum prediction error is given [19, p. 568] by

$$\begin{aligned} E_{\min}(\omega_{k-1}, \omega_k) &= \min_{\alpha_k,\beta_k} E(\omega_{k-1}, \omega_k|\alpha_k, \beta_k) \\ &= r_k(0) - \alpha_k^{\text{opt}} r_k(1) - \beta_k^{\text{opt}} r_k(2). \qquad (6) \end{aligned}$$

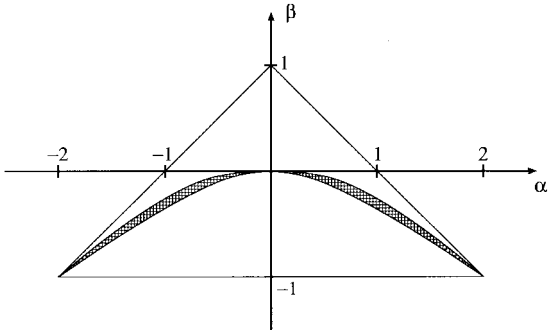


Fig. 2. Illustrations of the allowed regions for (α, β).

B. The Resonance Conditions

The resonance frequency we have derived is different from the usual definition of the resonance frequency in the context of second-order models. Also, there are constraints on the values that can be taken on by the prediction coefficients. Therefore, we discuss the relationships between these different approaches. Since we fix a specific segment k in the following, we drop the index k for this discussion and use the symbols α and β for the prediction coefficients.

From the minimum requirement, we obtain the constraint that the zeros of the complex predictor polynomial $A(z)$ with complex z must lie inside the unit circle [15, pp. 159–160]. Using these identities, the minimum requirement can be expressed in terms of the prediction coefficients α and β [3, p. 60], as follows:

$$\begin{aligned}\beta+\alpha&<1\\ \beta-\alpha&<1\\ |\beta|&<1.\end{aligned}$$

These constraints result in a triangular region in the (α, β)-plane, as shown in Fig. 2. This figure also contains other constraints, which are discussed next.

In order to model a pole, i.e., a true second-order resonator, the conventional approach is to require that the zeros of the predictor polynomial should form a conjugate complex pair [3, p. 60]. This requirement results in the constraint

$$\alpha^2+4\beta<0$$

which by combining with the previous constraints can be tightened to the new constraints

$$\begin{aligned}|\alpha|&<2\\ -1<\beta&<\frac{-\alpha^2}{4}.\end{aligned}$$

These constraints result in a parabolic boundary line as shown in Fig. 2. In the case of such a conjugate complex pair, the so-called pole frequency ϑ is given by the equation

$$\cos\vartheta=\frac{\alpha}{2\sqrt{(-\beta)}}.$$

In the approach presented in this paper, we define the relevant frequency as that frequency at which the magnitude of the predictor polynomial attains its minimum. As shown before, this resonance frequency φ is given by the equation

$$\cos\varphi=-\frac{\alpha(1-\beta)}{4\beta}.$$

From the evident inequality $|\cos\varphi|<1$, we obtain the following constraints for α and β:

$$\begin{aligned}|\alpha|&<2\\ -1<\beta&<\frac{-|\alpha|}{4-|\alpha|}.\end{aligned}$$

Plotting the corresponding boundary lines in the (α, β)-plane as shown in Fig. 2, we see that these constraints are tighter than the constraints for a pole solution. This relation is easy to prove. For $0<|\alpha|<2$, we can write down the following sequence of inequalities:

$$\begin{aligned}0<|\alpha|\cdot(|\alpha|-2)^2&=-\alpha^2\cdot(4-|\alpha|)+4|\alpha|\\ -4|\alpha|&<-\alpha^2\cdot(4-|\alpha|)\\ \frac{-|\alpha|}{4-|\alpha|}&<\frac{-\alpha^2}{4}.\end{aligned}$$

Thus, it can be seen that the resonance condition always implies the pole condition. The two frequencies converge to the same value if the damping of the pole approaches zero, which is given by $\beta\to(-1)$.

III. Dynamic Programming Algorithm for Segmentation

So far we have considered the prediction error of a single segment k only. We now assume that the whole frequency range is divided into K segments with boundaries $\omega_0=0,\cdots,\omega_{k-1},\omega_k,\cdots,\omega_K=\pi$. In this section, we describe a dynamic programming algorithm for finding the optimum segment boundaries.

To define the prediction error for the whole frequency range, we have to sum up the errors of all segments

$$E=\sum_{k=1}^{K}E_{\min}(\omega_{k-1},\omega_k).$$

In order to compute the autocorrelation coefficients $r_k(\nu)$, we use a discrete approximation of the integral in (5). The frequency interval $[0,\pi]$ is sampled at $I+1$ equally spaced frequencies $\pi i/I$, $i=0,1,\cdots,I$. The segment boundaries $\omega_0=0,\cdots,\omega_k,\cdots,\omega_K=\pi$ are replaced by the indices $i_0=-1,\cdots,i_k,\cdots,i_K=I$. The autocorrelation coefficients $r_k(\nu)$ are then given by

$$r_k(\nu)=\frac{1}{I}\sum_{i=i_{k-1}+1}^{i_k}|S(i)|^2\cos\left(\frac{2\pi\nu i}{2I}\right)$$

with

$$S(i):=S[e^{j(2\pi i/2I)}].$$

As usual, the discrete short-time power spectrum $|S(i)|^2$ is computed using a fast Fourier transform with $(2\cdot I)$ points. Because of the symmetry properties of the short-time power

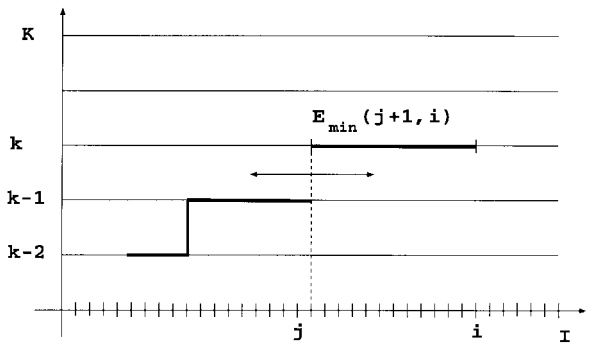


Fig. 3. Segmentation by dynamic programming.

spectrum, only the frequencies in the interval $[0, \cdots, I]$ are relevant. The autocorrelation coefficients can be computed efficiently using the identity

$$r_k(\nu) = T(\nu, i_k) - T(\nu, i_{k-1}) \tag{7}$$

with look-up tables

$$T(\nu, i) = \frac{1}{I} \sum_{i'=0}^{i} |S(i')|^2 \cos\left(\frac{2\pi\nu i'}{2I}\right) \tag{8}$$

for $\nu = 0, 1, 2$ and $i = 0, 1, \cdots, I$.

The task is now to find the segment boundaries $i_1, \cdots, i_{K-1}$ so that

$$\sum_{k=1}^{K} E_{\min}(i_{k-1} + 1, i_k)$$

is minimized. Dynamic programming [2], [4] provides an efficient solution. We introduce an auxiliary quantity $F(k, i)$, which is defined as the error of the best segmentation of the frequency interval $[0, i]$ into k segments. By decomposing the frequency interval $[0, i]$ into two frequency intervals, $[0, j]$ and $[j+1, i]$, and using the optimality in the definition of $F(k, i)$, we obtain the recurrence relation of dynamic programming

$$F(k, i) = \min_j [F(k-1, j) + E_{\min}(j+1, i)]. \tag{9}$$

As (9) shows, the best segmentation of the frequency interval $[0, j]$ into $k-1$ segments is utilized to determine the partition of the frequency interval $[0, i]$ into k segments. Fig. 3 gives an illustration of how (9) performs an optimum segmentation. To keep track of the optimum segmentation boundary used to compute $F(k, i)$, it is convenient to introduce associated backpointers, denoted as $B(k, i)$, that simply store the optimum boundary

$$B(k, i) = \arg\min_j [F(k-1, j) + E_{\min}(j+1, i)].$$

The optimum segment boundaries are obtained by recursively applying (9). The minimum overall error is then given by $F(K, I)$. Table I summarizes the complete algorithm. The first step is to fill the look-up tables defined by (8). Then

TABLE I
DYNAMIC PROGRAMMING ALGORITHM FOR FINDING THE SEGMENT BOUNDARIES

initialisation:
- compute $E_{min}(j, i)$ for $0 \le i \le I$, $0 \le j \le i$
- $F(0, i) = \infty$ for $0 \le i \le I$
- $F(k, -1) = \infty$ for $1 \le k \le K$
- $F(0, -1) = 0$

for each frequency $i = 0, 1, \ldots, I$ do
  for each segment $k = 1, 2, \ldots, \min(i+1, K)$ do
    $F(k, i) = \infty$
    for each frequency $j = k-2, k-1, \ldots, i-1$ do
      if $F(k-1, j) + E_{min}(j+1, i) < F(k, i)$
        $F(k, i) = F(k-1, j) + E_{min}(j+1, i)$
        $B(k, i) = j$

start traceback at $i(K) = I$

for each segment $k = K, K-1, \ldots, 1$ do
  $i(k-1) = B(k, i(k))$
  calculate formant parameters

the values of $E_{\min}(j, i)$ for $0 \le i \le I$ and $0 \le j \le i$ are calculated using (6) and (7). The algorithm employs the array B to store the backpointers and to construct the optimum set of segment boundaries. After this segmentation process, the formant frequencies for each segment are calculated by (3). This could be avoided by storing also the optimum formant parameters. However, this increased computational effort is negligible in comparison with the effort for finding the optimum segmentation. This effort is determined by the following two calculations:

1) the filling of the table $E_{\min}(i, j)$, which requires $I^2/2$ operations;
2) the dynamic programming recursion, which requires $K \cdot I^2/2$ operations.

We see that, in both cases, the computational effort is quadratic in the number of frequency lines I that are evaluated as segment boundaries in the dynamic programming recursion. In the following, we will distinguish between the number of frequency lines that are used for the approximation of the Fourier integral and the number of frequency lines that are hypothesized as segment boundaries in the dynamic programming optimization. Using this distinction, the complexity of the dynamic programming algorithm is then determined by the square of the number of segment boundary candidates. Therefore, the overall computational effort can be reduced significantly, if the segmentation evaluates only every mth frequency line of the discrete Fourier transform as possible segment boundary, where m is typically 2, 4, 8, 16. Nevertheless, for the estimation of the formant parameters, all frequency lines are used. In our implementation, the reduction of the frequency resolution for the segmentation is achieved by an additional step-size parameter in the loops over the frequency axis in the dynamic programming algorithm. In order to keep the computational effort at an affordable level, we have to check experimentally how much frequency resolution is really needed for a reliable determination of the segment boundaries.

IV. EXPERIMENTAL RESULTS

The formant model presented in the previous sections was tested in recognition experiments on the TI digit string data

base. This section reports on the results and is divided into two subsections. The first subsection gives several illustrating examples. In particular, we will show short-time spectra with estimated formant models, formant tracks superimposed on the time-frequency spectrograms and histograms of formant frequencies.

In the second subsection, we will focus our attention on formant-based speech recognition. The estimated formant frequencies are used to form the acoustic vectors for recognition experiments. In a series of experiments, the optimum acoustic vector was established. Using this optimum acoustic vector, further recognition experiments were conducted. In particular, we tested two types of density modeling, namely Gaussians and Laplacians, and we compared the error rates obtained with the mel-cepstrum and with formants. Furthermore, we investigated the effect of the spectral resolution on the recognition performance in order to study the trade-off between recognition performance and computational effort of the formant estimation.

A. Examples of Formant Estimates

In the following, we first give the details of our signal processing implementation and then discuss a variety of examples of formant estimation. In our experiments, we used the TI digit string data that had been sampled at a rate of 20 kHz. First, we perform a signal preemphasis by calculating the first-order difference of the sampled speech signal. Every 10 ms, a 20-ms Hamming window is applied to overlapping speech segments, and the short-time power spectrum is computed by a 1024-point fast Fourier transform (FFT). The frequency range from 0–5 kHz is used for formant estimation so that we have 256 samples in this range that are used for the approximation of the Fourier integral and for finding the optimum segmentation. The formant and bandwidth frequencies are determined by the model we have described; there is no smoothing of the formant frequencies. We do not adapt the frequency range to each speaker nor do we exclude the low-frequency range.

Next, we present experimental results of formant estimation in order to illustrate the properties of the proposed formant estimation algorithm. Figs. 4 and 5 show examples of formant models along with the corresponding short-time power density spectrum. There are two types of vertical lines for the formants, namely the formant frequencies and the segment boundaries. The values of the resonance frequencies are shown on the frequency axis, whereas the segment boundaries are represented as vertical lines without any numbers. In Fig. 4, we keep the number of formant models fixed at $K = 4$. Fig. 4(a) is an example of the "AY" vowel in "five." The figure shows the spectrum and the estimated formant models for frame 35 of the digit string *5873* spoken by male talker *IF*. The estimated formant frequencies are 756, 1270, 2422, and 3369 Hz. These values agree well with the synthesis parameters given in [1, p. 186]. Frame 95 of the utterance *73* by male talker *AH* is shown in Fig. 4(b). This frame is part of the transition between the R semivowel and the "IY" vowel in "three." Fig. 4(c) and (d) display further examples. These examples indicate that the proposed algorithm for formant estimation allows a reliable estimation of formant frequencies.

However, there are cases where the fixed number of formants leads to problems. This is illustrated by Fig. 5. The figure shows frame 132 of the string *554* spoken by male talker *HN*. Looking at the spectrum, we would expect $K = 5$ formants. In Fig. 5(a), however, only $K = 4$ formant models were estimated. As a consequence, the formant model with the resonance frequency 3683 Hz does not match the spectrum. If the number of formant models is chosen as $K = 5$, the estimated formant models result in a much better fit to the spectrum, as can be seen in Fig. 5(b).

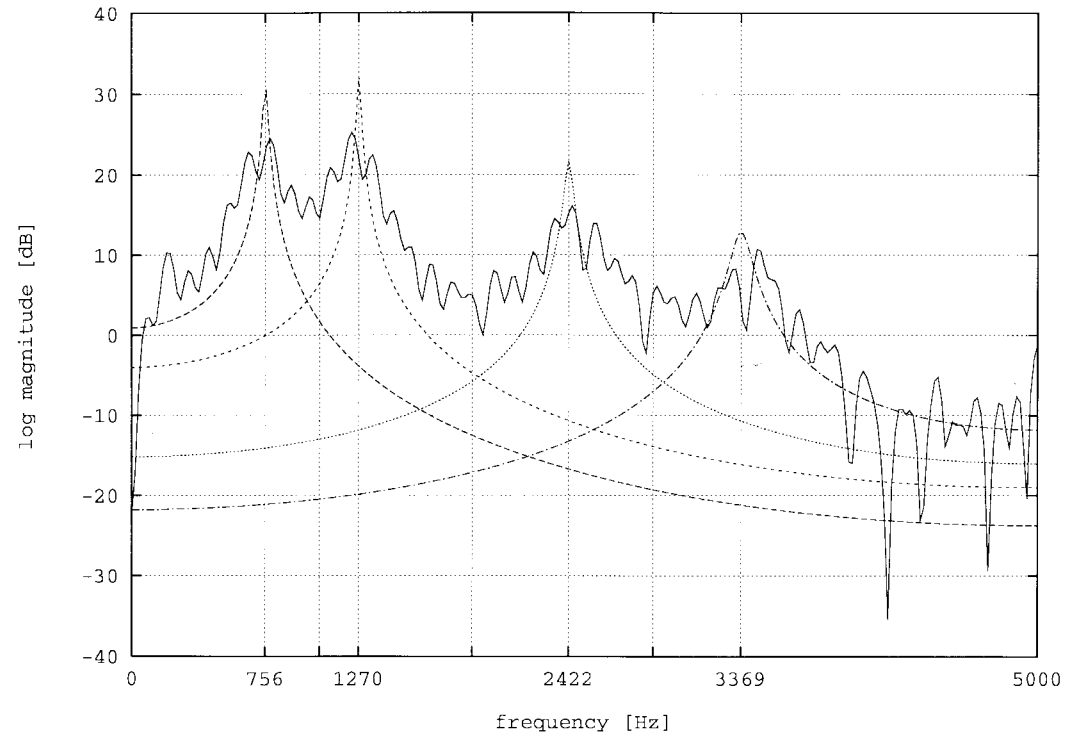
Fig. 6 presents examples of formant tracks superimposed on the spectrogram, i.e. the sequence of short-term power density spectra. Fig. 6(a) displays the formant contours for the digit string *5873* spoken by male talker *IF*, and Fig. 6(b) shows the string *94z* spoken by female talker *HG*. In this work and in [17], the character z represents "0" spoken as "zero." There is a good agreement between the formant frequencies and the spectrograms in regions of speech. As indicated in Fig. 6, the formant contours are significantly less smooth in areas of obvious silence. In the recognition experiments to be reported on later, we did not observe that the silence regions resulted in special problems for formant-based speech recognition. A possible explanation of this effect could be that silence portions are mainly recognized by making use of the frame energy and therefore the "jumpy" formant contours in silence regions are not critical.

Fig. 7 shows histograms for each of the four formant frequencies. The histograms were computed separately for male [Fig. 7(a)] and female speakers [Fig. 7(b)]. For the histograms, the silence frames in the acoustic signal were omitted. Evidently, these histograms depend on the spoken words and on the speaker population. Nevertheless, they are in reasonable agreement to the formant ranges reported in [1, p. 132], [10, p. 51], [5], and [22]. Comparing the formant frequencies in Fig. 7(a) and (b) also shows that on the average the formants of the female speakers are higher than the corresponding formants of the male speakers.

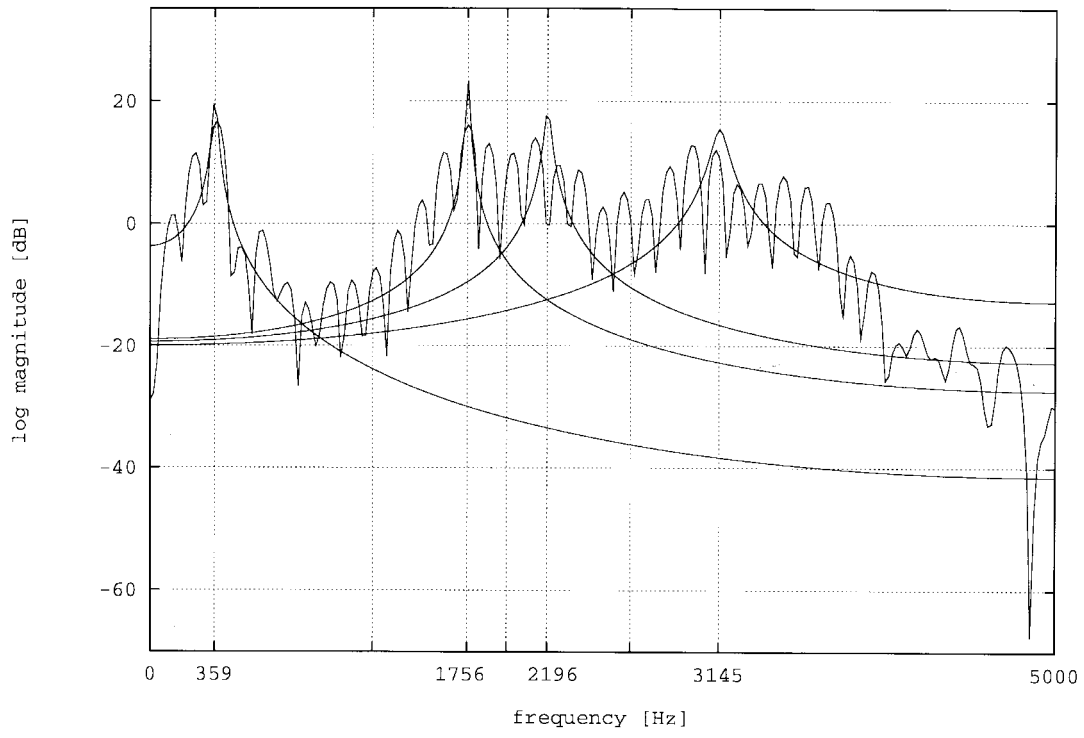
B. Formant-Based Speech Recognition

We will now present the results of a series of experiments that were conducted to optimize and study the performance of formant-based speech recognition. All recognition experiments were carried out on the adult portion of the TI digit string data base [18]. We used all available training and testing data. Thus, we had 8623 digit strings spoken by 55 male and 57 female speakers for training and 8700 digit strings uttered by 56 male and 57 female speakers for testing.

The recognition system [29] is based on whole-word hidden Markov models with continuous observation densities. The hidden Markov models are left-to-right models with forward, skip, and loop transitions. The system has gender-dependent word models for 11 English digits, including "*oh*" and gender-dependent silence models. The total number of states and, thus, emission distributions is 357 states plus one state for silence per gender. The emission probabilities are modeled as single



(a)



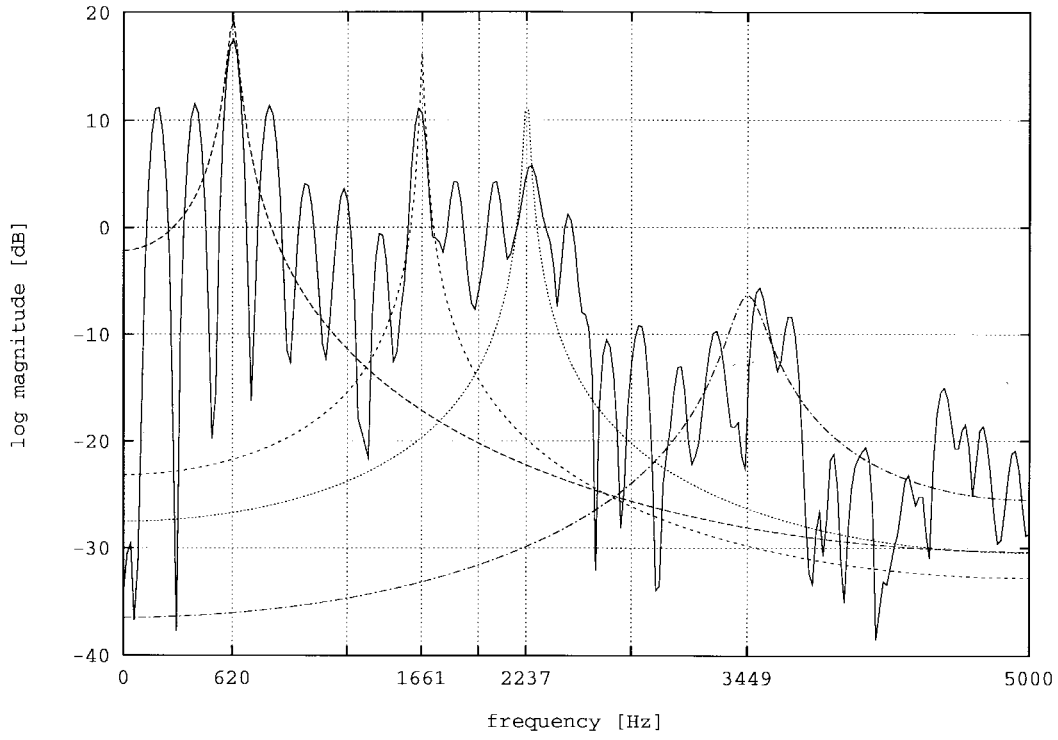
(b)

Fig. 4. Examples of formant models. (a) Frame 35, string *5873*, male talker *IF*. (b) Frame 95, string *73*, male talker *AH*.

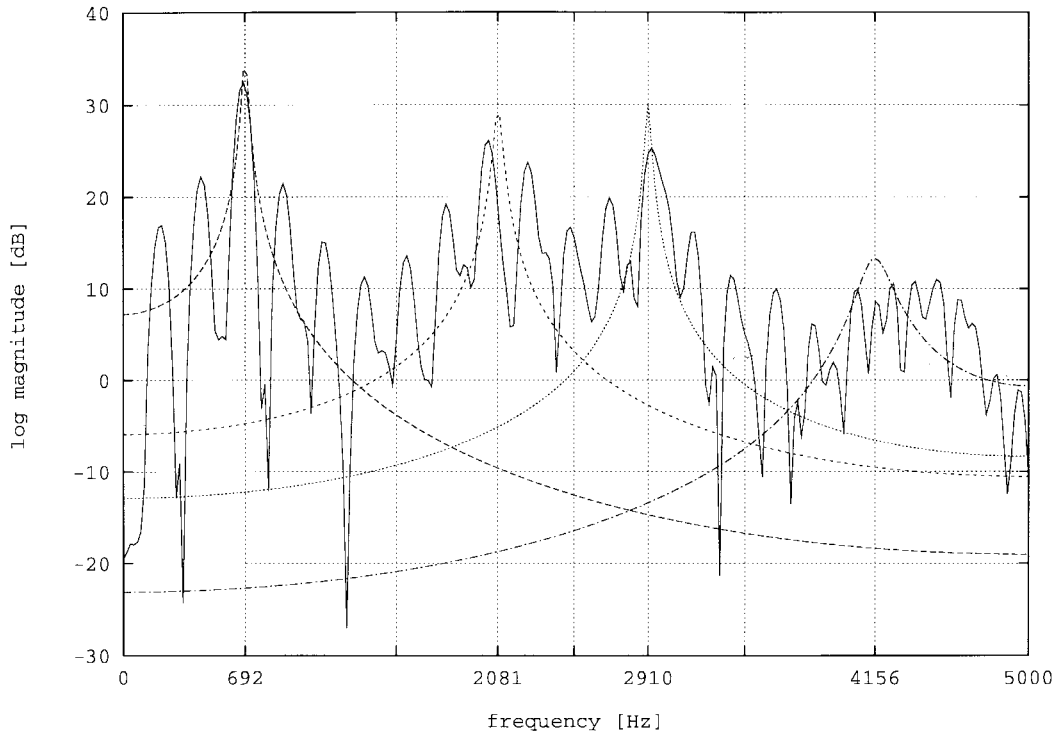
Laplacian densities with state-dependent deviation vectors. They are trained using the maximum likelihood criterion and the Viterbi approximation.

The recognition results are reported in terms of both string and word error rate. As usual, the word error rate is computed from the minimum number of deletion, substitution, and insertion errors. There were no syntactic constraints used in recognition, i.e., any sequence of digits was legal from the viewpoint of the recognizer.

The remainder of this subsection is divided into three parts. In the first part, we investigate how the formant parameters, i.e., resonance frequencies and bandwidths, can be used to



(c)

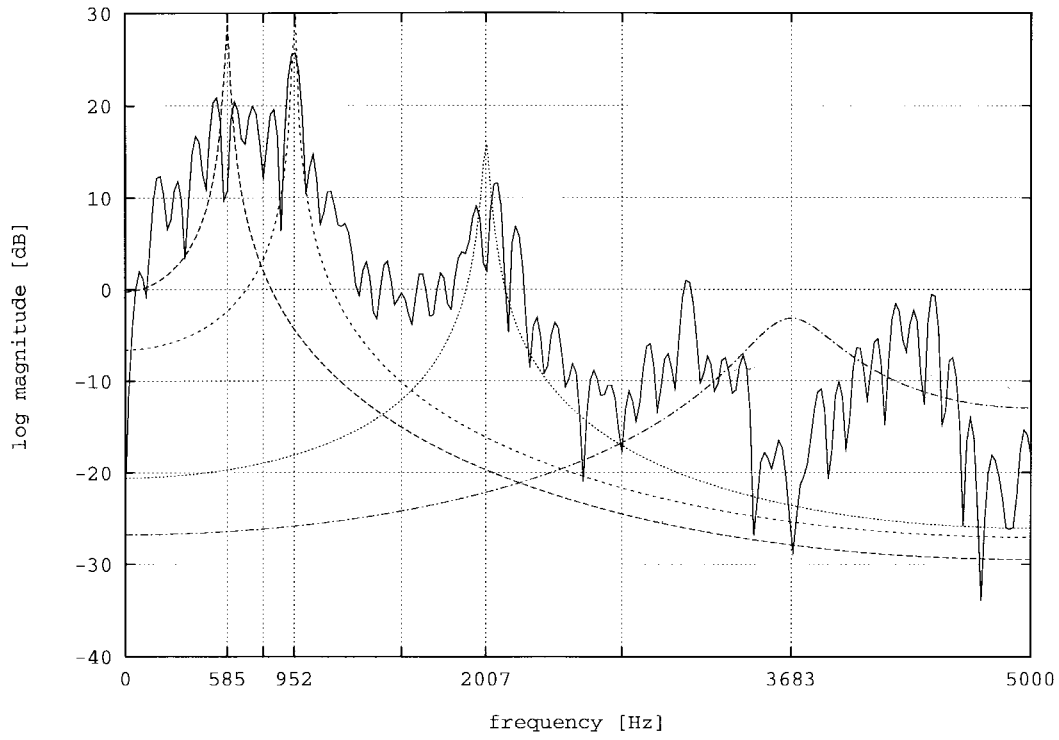


(d)

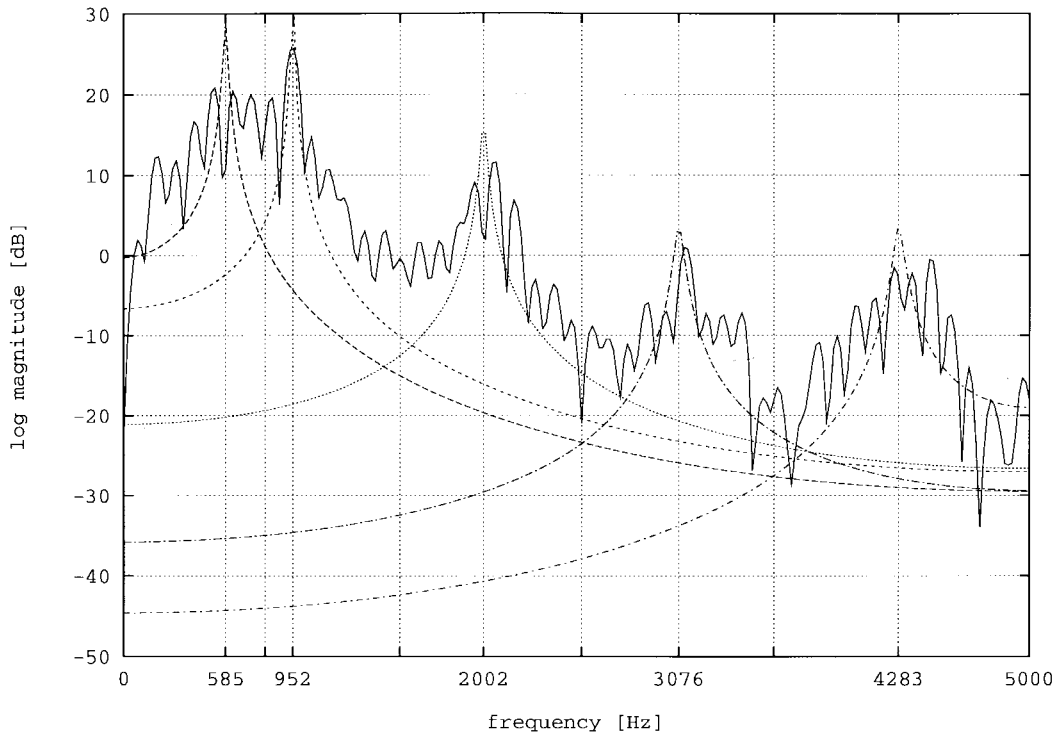
Fig. 4. (*Continued.*) Examples of formant models. (c) Frame 152, string *94z*, female talker *HG*. (d) Frame 67, string *74*, female talker *CM*.

form an acoustic vector for speech recognition. In the second part, the optimum acoustic vector is used to study the type of emission probability modeling and the acoustic resolution used in formant estimation. Also, recognition experiments comparing cepstral features and formant features are presented. A formant-based reference model is discussed in the third part.

1) Definition of the Acoustic Vector: In the following, we present experiments that were conducted to form the optimum acoustic vector for formant-based speech recognition. The experiments were started using only the formant frequencies along with the frame energy without the bandwidths. The acoustic vector was augmented by the first-order and second-



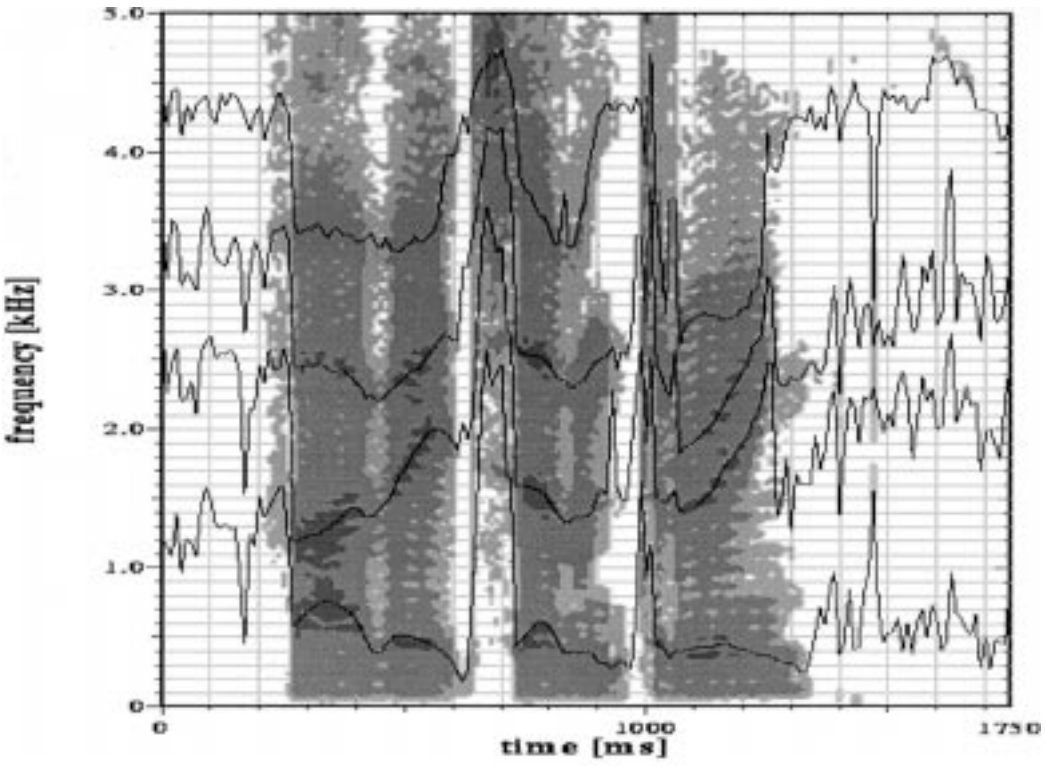
(a)



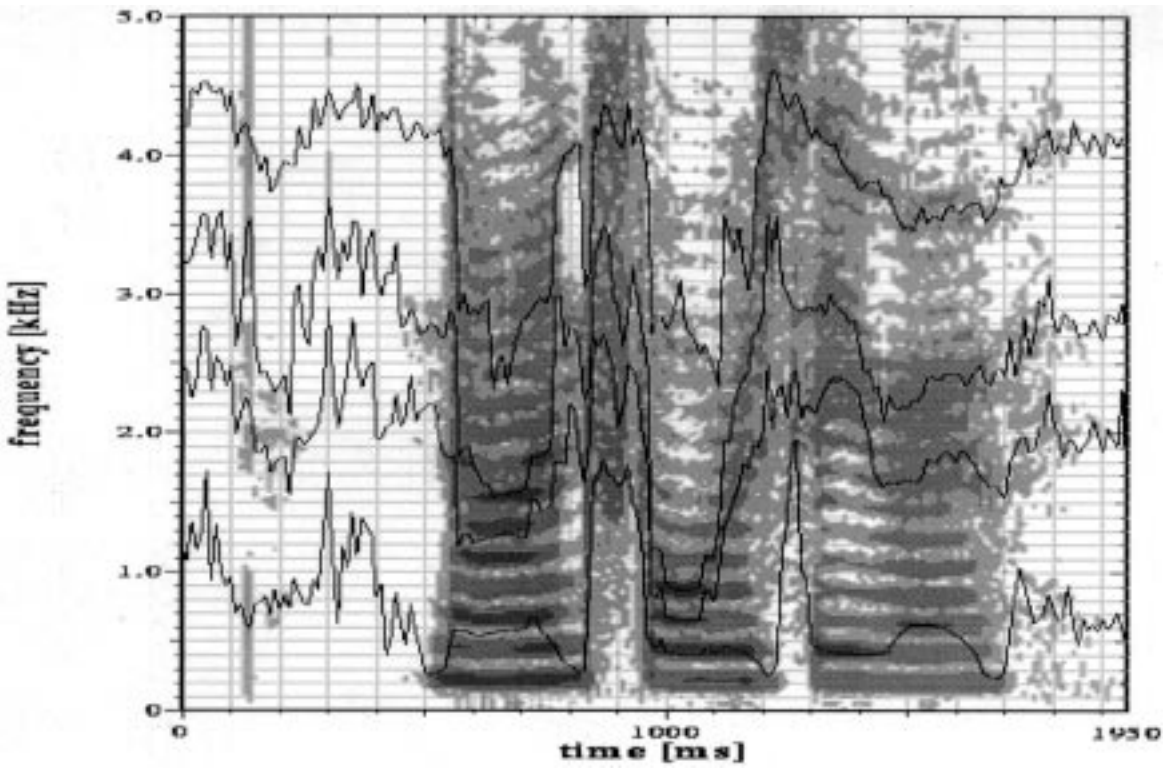
(b)

Fig. 5. Example of formant models for frame 132 of digit string *554* spoken by male talker *HN*. (a) Four formant models. (b) Five formant models.

order derivatives of the frame energy and by the first-order derivatives of the formant frequencies. For time frame t, the first-order derivatives were calculated from frames $t-3$ and t, and the second-order derivatives were calculated from frames $t-3$, t and $t+3$. The recognition results for these experiments are summarized in Table II. In Table II, two numbers for the formant frequencies are reported, namely one for estimation and one for recognition. The difference comes from the fact that, in some recognition experiments, the highest formant of the formants estimated was left out for recognition. It can be seen from Table II that the lowest error rate, both string error and word error rate, is obtained, if four formant



(a)

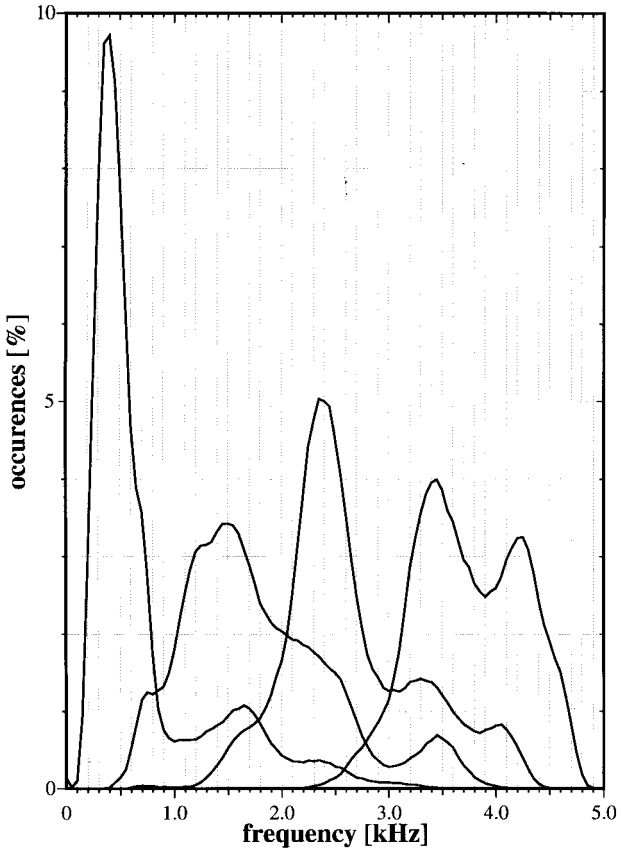


(b)

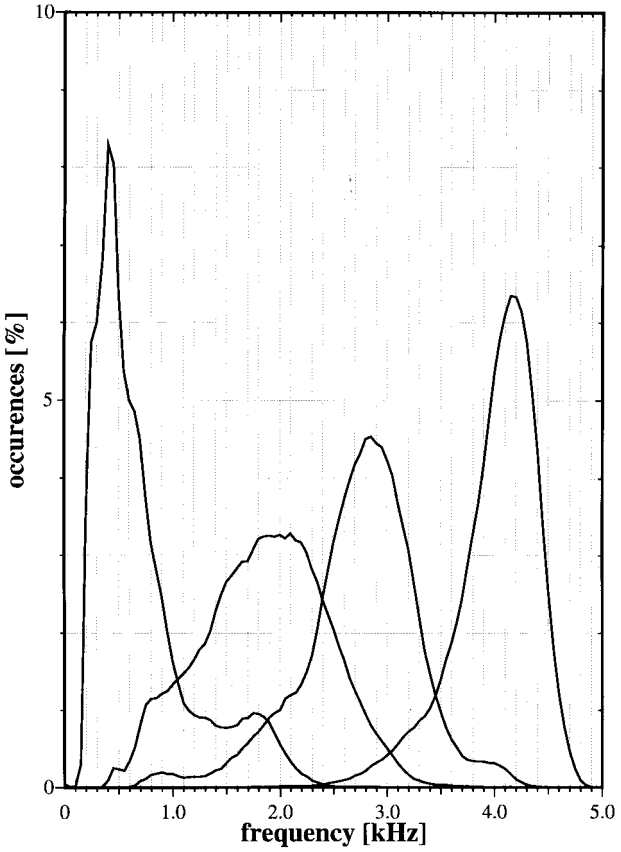
Fig. 6. Spectrogram and formant contours. (a) String *5873* by male talker *IF*. (b) String *94z* by female talker *HG*.

frequencies are estimated and the three lowest ones are used for recognition.

Next, we investigated the influence of the formant bandwidths on the recognition performance. The results are summarized in Table III. The formant estimation was based on $K = 4$ formant models. To form the acoustic vector, the bandwidths were used in a way similar to the formant frequencies, namely by selecting the three lowest formants and using the



(a)



(b)

Fig. 7. Histograms of formant frequencies over training speakers of TI digit string data base. (a) Male speakers. (b) Female speakers.

TABLE II
RECOGNITION RESULTS USING VARIOUS NUMBERS OF FORMANTS

Number of formants		Number of	Word error	String error
in estimation	in recognition	components	rate [%]	rate [%]
5	5	$2 \cdot 5 + 3$	1.9	5.4
	4	$2 \cdot 4 + 3$	2.0	5.6
4	4	$2 \cdot 4 + 3$	1.6	4.5
	3	$2 \cdot 3 + 3$	1.5	4.2
3	3	$2 \cdot 3 + 3$	3.0	8.1

TABLE III
EFFECT OF BANDWIDTH PARAMETERS ON THE ERROR RATE

Bandwidths	First-order derivatives of bandwidths	Number of components	Word error rate [%]	String error rate [%]
no	no	$2 \cdot 3 + 3$	1.5	4.2
yes	no	$2 \cdot 3 + 3 + 3$	1.7	4.8
	yes	$2 \cdot 6 + 3$	1.9	5.4

TABLE IV
EFFECT OF SECOND-ORDER DERIVATIVES ON THE ERROR RATES

Second-order derivatives of		Number of	Word error	String error
frame energy	formant frequencies	components	rate [%]	rate [%]
no	no	$2 \cdot 3 + 2$	1.7	4.9
yes	no	$2 \cdot 3 + 3$	1.5	4.2
	yes	$3 \cdot 3 + 3$	1.9	5.4

corresponding first-order derivatives. The results indicate that, for the given recognition conditions, the formant bandwidths have an adverse effect on the recognition results: The string error rate increases from 4.2% to 5.4% when the bandwidths and their first-order derivatives are included. Using only the bandwidths increases the string error rate from 4.2% to 4.8%. This suggests that any type of bandwidth information seems to deteriorate the recognition results.

Another experiment was carried out to check the usefulness of the second-order derivatives of the frame energy and of the formant frequencies. As the recognition results in Table IV show, omitting the second-order derivative of the frame energy increases the string error rate from 4.2% to 4.9%. For the formant frequencies themselves, there was no improvement when the second-order derivatives were added. In fact, as shown in Table IV, the string error rate goes up from from 4.2% to 5.4%.

As a result of the first phase of experiments, we had found that the optimum acoustic vector consists of

- the signal energy plus the corresponding first- and second-order derivatives;
- the three lowest formant frequencies from four estimated formants plus the corresponding first-order derivatives.

2) Recognition Results: The optimum acoustic vector was used to study the type of emission probability modeling and the acoustic resolution used in formant estimation. Also, a comparison of recognition results obtained with the mel-cepstrum and with formants was done. The results are summarized below.

Table V presents error rates for Gaussian and Laplacian densities. Replacing Laplacian densities by Gaussian densities increases the string error rate from 4.2% to 4.6%.

Table VI compares recognition results for formants and mel-cepstrum on the TI digit string data base. For both types of acoustic vectors, Table VI shows the word error rates, string

TABLE V
RECOGNITION RESULTS FOR GAUSSIAN AND LAPLACIAN DENSITIES

Density model	Number of components	Word error rate [%]	String error rate [%]
Laplacians	$2 \cdot 3 + 3$	1.5	4.2
Gaussians	$2 \cdot 3 + 3$	1.6	4.6

TABLE VI
COMPARISON OF RECOGNITION RESULTS FOR FORMANTS AND MEL-CEPSTRUM WITH SINGLE DENSITIES

Acoustic vector	Number of components	Word error rate [%]	String error rate [%]
Formants	$2 \cdot 3 + 3$	1.5	4.2
Mel-cepstrum:			
- full vector	$2 \cdot 15 + 3$	0.6	1.7
- reduced vector, with LDA	9	1.3	3.8
- " " , without LDA	$2 \cdot 3 + 3$	1.9	5.2

error rates, and the number of components of the acoustic vector. In the first experiment with the mel-cepstrum, the acoustic vector consists of 16 cepstrum coefficients, 16 first-order derivatives, and the second-order derivative of the signal energy [29]. The string error rate for the mel-cepstrum vector was then 1.7%. It should be emphasized that this error rate was obtained for single-density modeling. When using mixture densities, the string error rate can be reduced down to 0.71%. This error rate is actually the lowest reported [7], [9], [11], [12], [21].

Two more results for the mel-cepstrum with an acoustic vector consisting of nine components are given in Table VI. Using the first four mel-cepstrum coefficients, the four corresponding first-order derivatives, and one second-order derivative of the signal energy, the string error rate was 5.2%. A string error rate of 3.8% was obtained when we applied a linear discriminant analysis (LDA) [14] to a large input vector. This input vector consisted of three successive 33-component mel-cepstrum vectors $x(t-1)$, $x(t)$ and $x(t+1)$, which include derivatives. A 9×99 transformation matrix was used to reduce the dimension of the acoustic vector from 99 to nine components. We used one transformation matrix per gender. The classes to which the LDA is applied were chosen to be the states of the hidden Markov models. As can be seen in Table VI, using the same number of parameters the recognition results were sligthly better for the mel-cepstrum combined with a LDA than for the formants: The string error rates are 4.2% and 3.8% for the formant vector and the cepstrum vector, respectively.

Finally, we report on the effects of the frequency resolutions used for the calculation of the autocorrelation coefficients and for the segmentation process on the recognition performance. As pointed out above, these two frequency resolutions can be different. We carried out experiments with several configurations. The results are summarized in Table VII. Two conclusions can be drawn. First, Table VII indicates that the number of frequency samples provided by the discrete Fourier transform for the calculation of the autocorrelation coefficients is not critical: In the range from 128 to 2048 spectral lines, the recognition performance is not affected if the resolution for segmentation is high enough. The second conclusion concerns the frequency resolution for segmentation.

TABLE VII
EFFECT OF SPECTRAL RESOLUTION ON THE ERROR RATE; RESOLUTION IS GIVEN IN TERMS OF THE NUMBER OF FREQUENCY SAMPLES

Frequency samples for		Word error	String error
Fourier transform	segmentation	rate [%]	rate [%]
2048	256	1.5	4.2
256	256	1.5	4.2
	64	1.6	4.3
	16	2.4	6.6
128	128	1.5	4.2

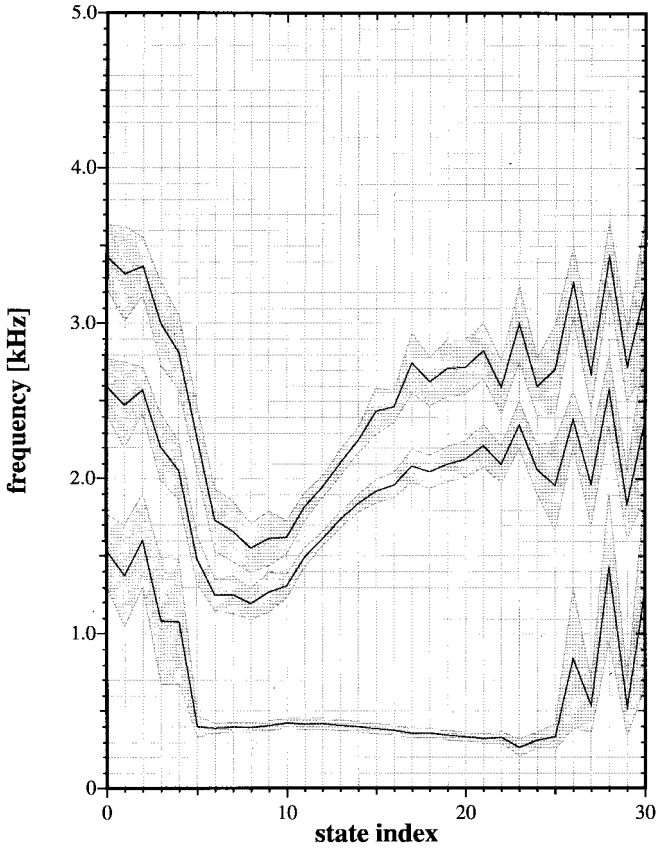
TABLE VIII
COMPARISON OF FORMANT FREQUENCIES OF THE VOWEL "IY" AND THE SEMIVOWEL "R." (a) FORMANT FREQUENCIES AVERAGED OVER FIVE STATES OF THE MALE REFERENCE MODEL FOR THE WORD "THREE." (b) FORMANT FREQUENCIES RECOMMENDED FOR SPEECH SYNTHESIS IN [1, P. 186]

		Formant frequencies [Hz]		
		F1	F2	F3
semivowel 'R'	a) this work	399	1252	1634
	b) Allen et al.	330	1060	1380
vowel 'IY'	a) this work	340	2114	2725
	b) Allen et al.	310	2200	2960

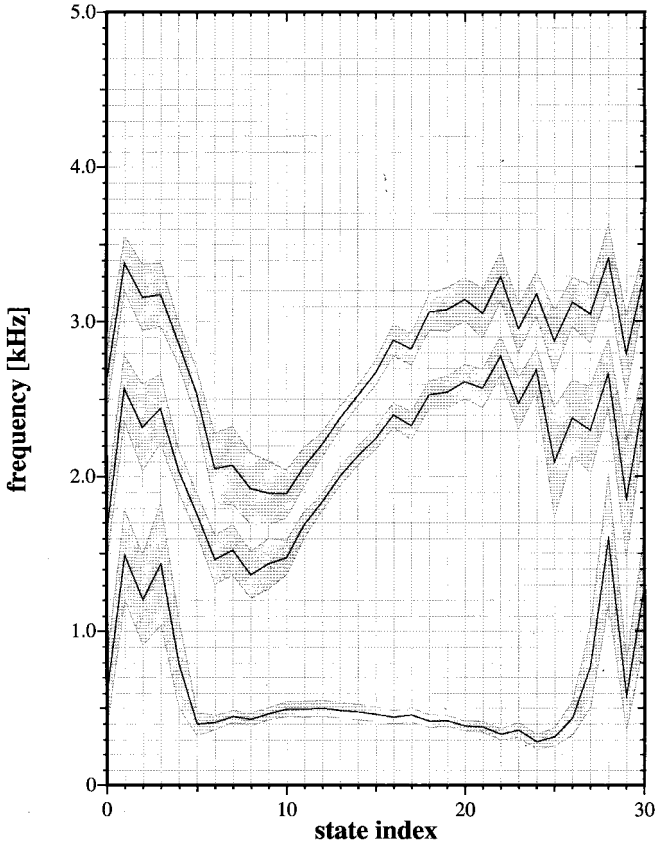
As Table VII shows, the error rate goes up significantly, if less than 64 segment boundaries are considered in the dynamic programming based segmentation.

3) Formant-Based Reference Models: For illustration purposes, we consider in more detail one specific reference model as it was obtained after training. Fig. 8 shows the male and female reference models for "three" using the optimum acoustic vector as defined above. In addition to the contours of the three formant frequencies, Fig. 8 shows the absolute deviation of the Laplacian models as a gray stripe around the formant frequency, which is estimated as the sample mean. To verify this reference, we selected by hand the formant frequencies for two sounds of the male reference model, namely "R" and "IY." For each of these two sounds, the formant frequencies were averaged over the associated states. For the "R" sound, we used states 6–10, and for the "IY" sound, we used states 17–21. The formant frequencies thus obtained are shown in Table VIII along with the formant frequencies that are typically used for speech synthesis and recommended in [1, p. 186]. As can be seen in Table VIII, there is a reasonable agreement between the two sets of formant frequencies. Similar results were obtained for the vowels in the reference models of the other digits. As usual in speech synthesis, the formant values for synthesis shown in Table VIII are based on the pole frequency definition. The differences between the pole frequency and the resonance frequency as used in this paper are negligible for formant frequencies above 1 kHz. For the first formant, however, the differences can be significant.

There is another observation that can be seen in Fig. 8. In a region ranging from state 22 to 30, the formant contours exhibit an oscillating behavior. To interpret this observation, we have to take into account that the hidden Markov models allow skip transitions. So the interpretation is that the oscillations reflect an implicit mixture modeling of the emission probabilities with two component densities. To verify this hypothesis, we checked the reference models using cepstral



(a)



(b)

Fig. 8. Formant-based reference models for the word "three." (a) Male model. (b) Female model.

acoustic vectors and observed the same effect. Therefore the conclusion is that these erroneous oscillations are *not* formant specific.

V. Conclusions

This paper has presented the following new approach to formant estimation.

1) The short-time power spectrum is decomposed into segments, each of which is modeled by a digital resonator.
2) The segment boundaries are optimized by dynamic programming.

The estimated formant frequencies have been analyzed using spectrograms and histograms. In a recognition test on the adult corpus of the TI digit string data base, a string error rate of 4.2% has been achieved with three formant frequencies and signal energy. A slightly better string error rate of 3.8% was obtained with the mel-cepstrum and the same number of parameters. In this experiment, a linear discriminant analysis was applied to reduce the dimension of the acoustic vector. There was no smoothing or other postprocessing of the formant trajectories. To the best of our knowledge, this is one of the few recognition systems that are based solely on formant contours. Considering that this work has started only recently, we see room for further improvements in formant-based speech recognition in the future. In this work, we have applied the formant estimation method to clean speech only. A topic for further research will be the application to noisy speech.

Acknowledgment

The authors thank S. Kanthak and V. Nussbaum for their assistance in providing the illustrations presented in this paper and in conducting the reported experiments.

References

[1] J. Allen, M. S. Hunnicutt, and D. H. Klatt, *From Text to Speech: The MITalk System.* Cambridge, MA: Cambridge Univ. Press, 1987.
[2] R. Bellman and S. Dreyfus, *Applied Dynamic Programming.* Princeton, NJ: Princeton Univ. Press, 1962.
[3] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis—Forecasting and Control*, 3rd ed. Englewood Cliffs, NJ: Prentice-Hall, 1994.
[4] J. S. Bridle and N. C. Sedgewick, "A method for segmenting acoustic patterns, with applications to automatic speech recognition," in *Proc. 1977 IEEE Int. Conf. Acoustics, Speech and Signal Processing,* Hartford, CT, May 1977, pp. 656–659.
[5] M. A. Bush and G. E. Kopec, "Network-based connected digit recognition," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-35, pp. 1401–1413, Oct. 1987.
[6] H. S. Chhatwal and A. G. Constantinides, "Speech spectral segmentation for spectral estimation and formant modeling," in *Proc. 1987 IEEE Int. Conf. Acoustics, Speech, and Signal Processing,* Dallas, TX, Apr. 1987, pp. 316–319.
[7] W. Chou, C.-H. Lee, and B.-H. Juang, "Minimum error rate training of inter-word context dependent acoustic model units in speech recognition," in *Proc. 1994 Int. Conf. on Spoken Language Processing,* Yokohama, Japan, Sept. 1994, pp. 439–442.
[8] A. Crowe and M. A. Jack, "Globally optimising formant tracker using generalized centroids," *Electron. Lett.,* vol. 23, pp. 1019–1020, Sept. 1987.
[9] G. R. Doddington, "Phonetically sensitive discriminants for improved speech recognition," in *Proc. 1989 IEEE Int. Conf. Acoustics, Speech, and Signal Processing,* Glasgow, U.K., May 1989, pp. 556–559.

[10] G. Fant, *Speech Sounds and Features.* Cambridge, MA: MIT Press, 1973.
[11] J. L. Gauvain and C. H. Lee, "Improved acoustic modeling with Bayesian learning," in *Proc. 1992 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, San Francisco, CA, Mar. 1992, vol. I, pp. 481–484.
[12] R. Haeb-Umbach, D. Geller, and H. Ney, "Improvements in connected digit recognition using linear discriminant analysis and mixture densities," in *Proc. 1993 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Minneapolis, MN, Mar. 1993, vol. II, pp. 239–242.
[13] M. J. Hunt and C. Lefèbvre, "Speaker dependent and independent speech recognition experiments with an auditory model," in *Proc. 1988 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, New York, Apr. 1988, pp. 215–218.
[14] ——, "A comparison of several acoustic representations for speech recognition with degraded and undegraded speech," in *Proc. 1989 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Glasgow, U.K., May 1989, pp. 262–265.
[15] S. M. Kay, *Modern Spectral Analysis—Theory and Application.* Englewood Cliffs, NJ: Prentice-Hall, 1988.
[16] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Amer.*, vol. 67, pp. 970–995, Mar. 1980.
[17] G. E. Kopec, "Formant tracking using hidden Markov models and vector quantization," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 709–729, Aug. 1986.
[18] R. G. Leonard, "A database for speaker-independent digit recognition," in *Proc. 1984 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, San Diego, CA, Mar. 1984, pp. 42.11.1–42.11.4.
[19] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.
[20] S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 135–141, 1974.
[21] Y. Normandin, R. Cardin, and R. De Mori, "High-performance connected digit recognition using maximum mutual information estimation," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 299–311, Apr. 1994.
[22] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *J. Acoust. Soc. Amer.*, vol. 24, pp. 175–194, Mar. 1952.
[23] L. Rabiner and R.-W. Schafer, *Digital Processing of Speech Signals.* Englewood Cliffs, NJ: Prentice-Hall, 1978.
[24] H. B. Richards, J. S. Mason, M. J. Hunt, and J. S. Bridle, "Deriving articulatory representations of speech," in *Proc. 1995 Europ. Conf. Speech Communication and Technology*, Madrid, Spain, Sept. 1995, pp. 761–764.
[25] O. Schmidbauer, "An algorithm for automatic formant extraction in continuous speech," in *Proc. EUSIPCO-90, Fifth European Signal Processing Conf.: Signal Processing V, Theories and Applications*, Barcelona, Spain, Sept. 1990, vol. 2, pp. 1151–1154.
[26] R. C. Snell and F. Milinazzo, "Formant location from LPC analysis data," *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 129–134, Apr. 1993.
[27] D. Talkin, "Speech formant trajectory estimation using dynamic programming with modulated transition costs," AT&T Int. Memo. MH 11222 2924, AT&T Bell Labs., Murray Hill, NJ, 1987.
[28] L. Welling and H. Ney, "A model for efficient formant estimation," in *Proc. 1996 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Atlanta, GA, May 1996, vol. 2, pp. 797–800.
[29] L. Welling, H. Ney, A. Eiden, and C. Forbrig, "Connected digit recognition using statistical template matching," in *Proc. 1995 Europ. Conf. Speech Communication and Technology*, Madrid, Spain, Sept. 1995, vol. 2, pp. 1483–1486.

Lutz Welling was born in Wuppertal, Germany, in 1968. He received the Dipl. degree in electrical engineering from Aachen University of Technology, Aachen, Germany, in 1993.

Since 1993, he has been with the Department of Computer Science, Aachen University of Technology. In 1996, he was a Visiting Researcher at the Entropic Research Laboratory, Washington, DC. His research interests are in large vocabulary speech recognition, robust speech recognition, and signal processing.

Hermann Ney (M'86) received the Dipl. degree in physics from the University of Goettingen, Goettingen, Germany, in 1977, and the Dr.-Ing. degree in electrical engineering from Braunschweig University of Technology, Braunschweig, Germany, in 1982.

In 1977, he joined Philips Research Laboratories, first in Hamburg, then in Aachen, Germany, where he worked on various aspects of speaker verification, isolated and connected word recognition, and large-vocabulary continuous-speech recognition. In 1985, he was appointed Head of the Speech and Pattern Recognition Group. From 1988 to 1989, he was a Visiting Scientist at AT&T Bell Laboratories, Murray Hill, NJ. In July 1993, he joined Aachen University of Technology, Aachen, Germany, as a Professor of computer science. His work has focused on the application of dynamic programming and statistical techniques for decision making in context. His current interests cover all aspects of pattern and speech recognition, such as signal processing, search strategies, language modeling, automatic learning, and language translation.