Gender recognition from speech. Part I: Coarse analysis

Ke Wu
Entropic Speech, Inc., 10011 North Foothill Boulevard, Cupertino, California 95014

D. G. Childers
Department of Electrical Engineering, University of Florida, Gainesville, Florida 32611-2024

(Received 18 July 1990; revised 23 February 1991; accepted 4 April 1991)

The purpose of this research was to investigate the potential effectiveness of digital speech processing and pattern recognition techniques for the automatic recognition of gender from speech segments. In this paper "coarse" acoustic coefficients (autocorrelation, linear prediction, cepstrum, and reflection) were used to form test and reference templates for vowels, voiced fricatives, and unvoiced fricatives. The effects of different distance measures, filter orders, recognition schemes, and vowels and fricatives were comparatively assessed to determine their effectiveness for the task of gender recognition from speech segments. The results showed that most of the acoustic parameters worked well for gender recognition. A within-gender and within-subject averaging technique was important for generating appropriate test and reference templates. The Euclidean distance measure appeared to be the most robust as well as the simplest of the distance measures. The results from this study implied that the gender information is time invariant, phoneme independent, and speaker independent for a given gender. One recognition scheme achieved 100% correct speaker gender classification for a database of 52 talkers (27 male and 25 female). In part II of this paper [D. G. Childers and K. Wu, J. Acoust. Soc. Am. **90**, 1841–1856 (1991); hereafter referred to as paper II] the detailed features of ten vowels that appeared responsible for distinguishing a speaker's gender were examined statistically. Included in paper II is a replication of part of the classical study of Peterson and Barney [J. Acoust. Soc. Am. **24**, 175–184 (1952)] of vowel characteristics.

PACS numbers: 43.70.Gr, 43.71.An, 43.71.Bp

INTRODUCTION

Human listeners appear capable of extracting information from the acoustic signal beyond just the linguistic message. Listeners are generally able to identify clues about the speaker's personality, emotional state, gender, age, dialect, accent, and the status of his/her health. Current automatic speech and speaker recognition systems are far less capable than human listeners. Computerized speech recognition and speaker verification can be accomplished but only under highly constrained conditions. Factors that limit automated speech and speaker recognition systems include our inability to identify acoustic features sensitive to the task and yet robust enough to accommodate speaker articulation differences, vocal tract differences, phonemic substitutions or deletions, prosodic variations, and other factors that influence our recognition ability. More insight and systematic study of intrinsically effective speaker discrimination features are needed. A series of small experiments should be done so that the experimental results will be mutually supportive and will lead to overall understanding of the combined effects of all the parameters that are likely to be present in actual situations (Rosenberg, 1976; Committee on Evaluation of Sound Spectrograms, 1979). One such problem is gender recognition or classification using acoustic features extracted from speech. An automatic gender recognition technique could assist the development of speaker-independent speech recognition systems, help identify acoustic features important for synthesizing male and female voices, and provide guidelines for identifying acoustic features related to dialect, accent, age, health, and other speaker idiosyncratic characteristics (Childers *et al.*, 1987, 1988).

The differences between male and female voices depend upon many factors. Generally, there exist three types of parameters: physiological and acoustical, which can be measured objectively, and perceptual, which is subjective but can be assessed psychophysically. Many physiological parameters of the male and female vocal apparatus have been determined and compared. Fant (1976) showed that the ratio of the total length of the female vocal tract to that of a male is about 0.87, and Hirano *et al.* (1983) showed that the ratio of the length of the female vocal fold to that of the male is about 0.8. Titze (1987, 1989) reported that, anatomically, the female larynx also differs from the male larynx in thickness, angle of the thyroid laminae, resting angle of the glottis, vertical convergence angle in the glottis, and in other ways.

The differences in physiological parameters can lead to differences in acoustical parameters. When comparing male and female formant patterns, the average female formant frequencies are roughly related to those of the male by a scaling factor that is inversely proportional to the overall vocal tract length. On the average, the female formant pattern is said to be scaled upward in frequency by about 20% compared to the average male formant pattern. Peterson and Barney (1952) measured the first three formant frequencies present in ten vowels spoken by men, women, and children. They reported that male formants were the lowest in fre-

quency, women had a higher range, and children had the highest. Schwartz (1968) also demonstrated that the gender of an individual can be identified from voiceless fricative phonemes such as [s] and [f].

The higher speaking fundamental frequency (pitch) range of the female speaker is quite well known. There is general agreement that the fundamental frequency is an important factor in the identification of gender from voice (Carlson 1981; Hollien and Malcik, 1967; Saxman and Burk, 1967; Hollien and Paul, 1969; Hollien and Jackson, 1973; Monsen and Engebretson, 1977; Stoicheff, 1981; Horri and Ryan, 1981; Linville and Fisher, 1985; Henton, 1987). One often finds the statement that the pitch level of the female speaking voice is approximately one octave higher than that of the male speaking voice (Linke, 1973). Titze (1989) found that the fundamental frequency was scaled primarily according to the membranous lengths of the vocal folds (scale factor 1.6). However, there is considerable variation among values obtained by different investigators. According to Hollien and Shipp (1972), the male subjects showed an intersubject pitch range of 112–146 Hz. Stoicheff's (1981) data showed that the range for the female subjects was 170–275 Hz.

During the last few years, measuring the area of the glottis as well as estimating the glottal volume–velocity waveform have become research topics of interest (Cheng and Guerin, 1987; Holmberg *et al.*, 1988). It is well known that the shape of the glottal excitation wave is an important factor that can affect speech quality (Rothenberg, 1981; Childers and Wu, 1990). The wave shape produced by male subjects is typically asymmetrical and frequently shows a prominent hump in the opening phase of the wave (due to source–tract interaction), while the female waveform tends to be symmetric with no hump during the opening phase, indicating less or no source–tract interaction (Fant, 1979). The closing portion of the wave generally occupies 20%–40% of the total period and there may or may not be an easily identifiable closed period (Monsen and Engebretson, 1977). Holmberg *et al.* (1988) found statistically significant differences in male–female glottal waveform parameters. In normal and loud voices, female waveforms indicated lower vocal fold closing velocity, lower ac flow, and a proportionally shorter closed phase of the cycle, suggesting a steeper spectral slope for females. For softly spoken voices, spectral slopes are more similar to those of males.

The perceptual parameters or strategies used to make decisions concerning male/female voices are not delineated in the literature, although making this discrimination appears to be performed routinely by human listeners (O'Kane, 1987). It is hypothesized that a limited number of perceptual cues for classifying voices can be identified by listeners, and these cues may include some sociological factors such as cultural stereotyping.

Singh and Murry (1978) and Murry and Singh (1980) investigated the perceptual parameters of normal male and female voices. They found that the fundamental frequency and formant structure of the speaker appeared to carry information that listeners used to assess voices. For example, effort, pitch, and nasality were the perceptual parameters used to characterize female voices while male voices were judged on the basis of effort, pitch, and hoarseness. The authors suggested that listeners may use different perceptual strategies to classify male voices than they use to classify female ones. Coleman (1976) also suggested that there was a possibility of a gender-specific listener bias for one acoustic characteristic or for one gender over the other.

Many researchers also believe melodic (intonation, stress, and/or coarticulation) cues are speech characteristics associated with female voices. Furthermore, the female voice is typically more breathy than the male voice (Klatt, 1987; Klatt and Klatt, 1990).

In summary, despite the fact that considerable knowledge is available in the literature about physiological, anatomical, and acoustical differences between male and female voice characteristics, only one attempt has been made to automatically classify male/female voices by objective feature measurements (Childers *et al.*, 1987, 1988). Previous research has used subjective listening tests to discriminate gender from speech. However, the results of the listening tests may depend on the gender distribution of the listening panel because males and females may use different judging strategies. Consequently, the conclusions reached from listening tests may be biased (Coleman, 1976; Carlson, 1981). Objective gender recognition techniques may assist other speech and speaker recognition procedures, as mentioned above.

The purpose of this study was to determine the effectiveness of digital speech processing and pattern recognition techniques for eventually developing an automatic gender recognition system. Emphasis was placed on the investigation of various objective acoustic parameters and distance measures, which in combination would be most effective for classifying a speaker's gender. The database and the techniques associated with data collection and preprocessing are discussed along with the various acoustic parameters and distance measures that were considered. The details of the template formation and recognition schemes are provided. The recognition performance based on "coarse" analysis of the acoustic features is described. Results of comparative studies of acoustic features, distance measures, recognition schemes, and filter orders are reported. The acoustic features were extracted from vowels, voiced fricatives, and unvoiced fricatives.

I. RESEARCH DESIGN

A. Database

Speech and electroglottographic (EGG) data were collected simultaneously from 52 talkers (27 male and 25 female) whose ages ranged from 20–80 years old. The data collection conditions were the following.

(1) Each subject was seated in an IAC single wall sound booth.

(2) An Electro Voice RE-10 dynamic cordiod microphone was located 6 in. from the speaker's lips.

(3) The electroglottograph was a Synchrovoice, Inc. model.

(4) The speech and EGG data were amplified prior to

digitization by a Digital Sound Corp. DSC-240 audio control console.

(5) The two data channels were directly digitized at a sampling frequency of 10 kHz per channel by a Digital Sound Corp. system with 16 bits precision. Both channels were bandlimited to 5 kHz by passive elliptic filters with a stopband attenuation of -55 dB and a passband ripple of ± 0.2 dB.

(6) The complete speech protocol consisted of 27 tasks, but this study was concerned only with ten sustained vowels, which for typing convenience we denote as /IY, I, E, AE, A, OW, U, OO, UH, ER/, five sustained unvoiced fricatives /H, F, THE, S, SH/, and four sustained voiced fricatives /V, TH, Z, ZH/. This notation is adopted from Rabiner and Schafer (1978). The subjects were instructed to pronounce (and sustain) each vowel as it would be pronounced in the following words, respectively: beet, bit, bet, bat, Bob, bought, book, boot, but, Bert. Similar instructions were given for the fricatives for which the cue words for the unvoiced and voiced fricatives, respectively, were hat, fix, thick, sat, ship, van, this, zoo, azure.

(7) The duration of each vowel and fricative approximated 2 s.

A detailed record was not obtained of each speaker's dialect. However, we did record the region, state, or country in which each speaker was raised. Five speakers were born in foreign countries but were raised in the U.S. from age 3–5. The majority of speakers (17) were raised in Florida; next was Ohio (5), Pennsylvania (4), Illinois (3), with others from the west coast, midwest, and the east coast. American English was the native language of all speakers except two; these two speakers were bilingual with no perceptible accent when they spoke English. A few of the U.S. born speakers were also bilingual having learned both Spanish and American English simultaneously; no accent was perceptible. Two colleagues in the University of Florida, Department of Speech, assisted in the data collection and describe the overall speaker's dialect as General American.

B. Analysis

Only the acoustic data were analyzed in this study. Asynchronous linear predictive coding (LPC) analysis was used because of the following.

(1) LPC features represent characteristics of the vocal source (except for fundamental frequency) as well as the vocal tract (Rabiner and Schafer, 1978).

(2) Formant frequency measurements obtained by LPC have been found to compare favorably to measures obtained by spectrographic analysis (Monsen and Engebretson, 1983; Linville and Fisher, 1985).

(3) The parameters used in gender recognition should be the same, if possible, as those used in speech or speaker recognition systems. The LPC model has been successfully applied in speech and speaker recognition (Makhoul, 1975; Atal, 1974, 1976; Rosenberg, 1976: Markel *et al.*, 1977; Davis and Mermerlstein, 1980; Rabiner and Levinson, 1981). Moreover, many related distortion or distance measurements have been developed (Gray and Markel, 1976; Gray *et al.*, 1980; Juang, 1984; Nocerino *et al.*, 1985), which could be conveniently adopted for the preliminary experiments of gender recognition.

(4) One can derive acoustic parameters from the LPC model using fast algorithms and short data records.

The LPC analysis used a fixed frame size, fixed frame rate and a fixed number of parameters per frame. The analysis conditions included (i) order of the filter: 8, 12, 16, 20; (ii) analysis frame size: 256 points/frame; (iii) frame overlap: None; (iv) preemphasis factor: 0.95; and (v) analysis window: Hamming.

The LPC coefficient calculations used a total of six frames from the data record for each utterance. The first two frames were selected from near the voice onset of the utterance, the second two frames from the middle of the utterance, and the last two frames from near the voice offset of the utterance. The six sets of coefficients obtained from these six frames were averaged to give a template coefficient for each sustained utterance. The reason for this choice of data frame locations was to make the templates more robust so that we might determine their usefulness for use with continuous speech.

A possible automatic gender recognition system is shown in Fig. 1. Aspects of this figure were used in our pilot study (Childers *et al.*, 1987, 1988) and were used in this study.

One of the key issues in developing a recognition system is to identify appropriate feature vectors and distance measures for calculating the separation between feature vectors that will support good recognition performance. Several acoustic parameters were considered as feature vector candidates in this study: autocorrelation coefficients (ARC), LPC, cepstrum coefficients (CC), and reflection coefficients (RC); all are described in Rabiner and Schafer (1978) and Furui (1989).

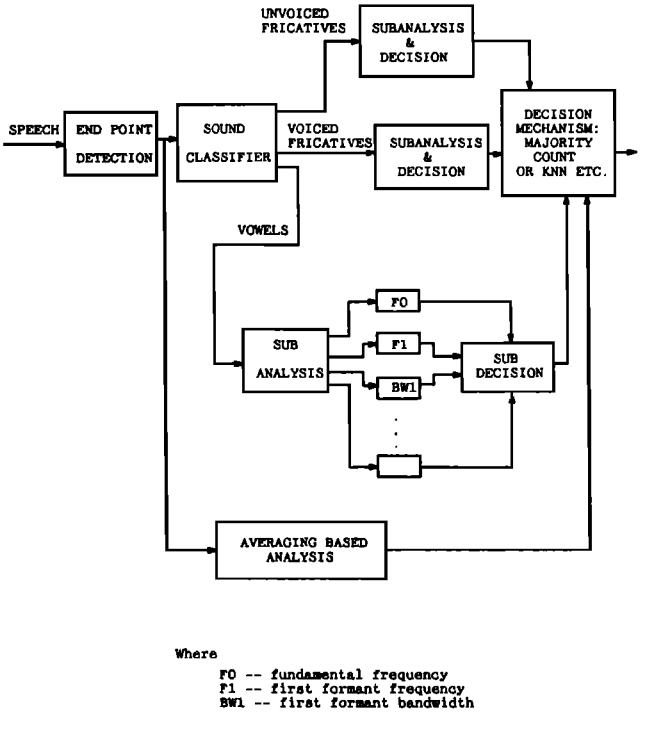

FIG. 1. A possible automatic gender recognition system.

In addition we considered fundamental frequency and formant (FFF) information. This set of features consisted of frequencies, bandwidths, and amplitudes of the first, second, third, and fourth formants and the fundamental frequencies of vowels (but not of fricatives). Formant information was obtained by a peak-picking technique, using an FFT of the LPC coefficients. Fundamental frequency was calculated using a modified cepstral algorithm.

The several distance measures we considered in the pilot study and considered here for assessing the separation between test and reference feature vectors are the Euclidean distance (EUC) (Furui, 1989), the LPC log likelihood distance (LLD) (Itakura, 1975), the cepstral distance (CD) (Nocerino *et al.*, 1985), the weighted Euclidean distance (WEUC) (Furui, 1989), and the probability density function (PDF), which is given as

$$D_{\mathrm{PDF}} = \left(\frac{1}{2\pi}\right)^{n/2} \frac{1}{|W|^{1/2}} \times \exp\left(\frac{-(X-Y)'W^{-1}(X-Y)}{2}\right), \quad (1)$$

where X is the test vector, W is the symmetrical covariance matrix obtained using a set of reference vectors (e.g., from a set of templates, which represent subjects of the same gender), and Y is the mean vector of this set of reference vectors. The weighting W compensates for correlation between features in the overall distance and reduces the intragroup variations. The weighted Euclidean distance is a simplified version of the likelihood distance measure using the probability density function. This distance measure minimizes the probability of error. Some authors refer to these distance measures as distortion measures. The decision rule for deciding the speaker and the speaker's gender was the minimum distance between the test and reference templates [the nearest neighbor (NN) rule]. The leave-one-out or jacknife procedure (Lachenbruch, 1975) was used to statistically evaluate the performance of the various feature vectors and distance measures.

II. TEMPLATE FORMATION AND RECOGNITION SCHEMES

A. Purpose

Two considerations in developing templates for speech pattern recognition are whether or not the data is time invariant and whether or not the task is text dependent or text independent. Several studies have considered averaging techniques to form feature vectors for speaker recognition that were text independent (Pruzansky, 1963; Markel *et al.*, 1977; Furui, 1989). Generally, these studies showed that the speaker recognition error rate improved or remained unchanged with data averaging, implying that speaker recognition could be achieved with a text-independent task. Temporal cues also appeared not to play a role in speaker gender identification (Lass and Mertz, 1978). Gender identification accuracy remained high and unaffected by temporal speech alterations such as playing speech samples backward or time compressing the speech samples.

Therefore, we hypothesize that gender information in speech is time invariant. Long-term data averaging should emphasize the speaker's gender information and increase the between-to-within gender variation ratio. In practice we should be able to achieve text-independent gender recognition prior to speech recognition or speaker verification, thereby potentially reducing the search space to as much as one-half.

The purpose for considering various test and reference template formation procedures was to verify the above hypothesis. Should the hypothesis prove correct we hoped to determine the amount of averaging necessary to accomplish gender recognition using only vowels, unvoiced fricatives, or voiced fricatives.

B. Test and reference template formation

The averaging procedures used to create test and reference templates for the present experiment employed a multilevel signal averaging approach as illustrated in some detail in Fig. 2. The token level templates were feature parameter vectors obtained from each utterance by an LPC analysis as described in the last section. The vectors were autocorrelation (ARC), LPC, cepstrum (CC), or reflection coefficients (RC). A token template was calculated by averaging six vectors obtained from six different frames for each sustained utterance (vowel, unvoiced fricative, or voiced fricative) for each speaker. The templates from each group of phoneme utterances from the token level for each speaker were averaged to form a speaker level template for each speaker. Thus

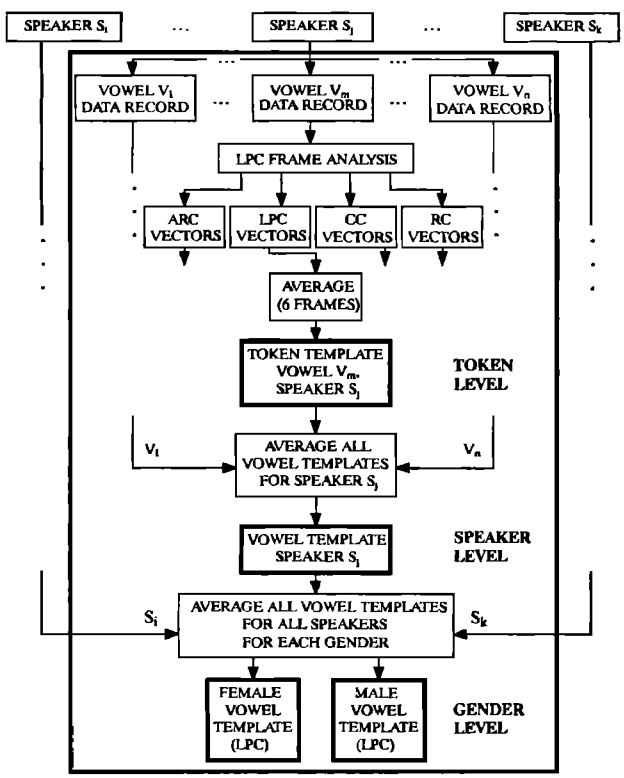

FIG. 2. Detail of template formation for vowels for LPC vectors at the token, speaker, and gender levels. In a similar manner templates were developed for ARC, CC, and RC vectors. The entire process was repeated for voiced and unvoiced fricatives.

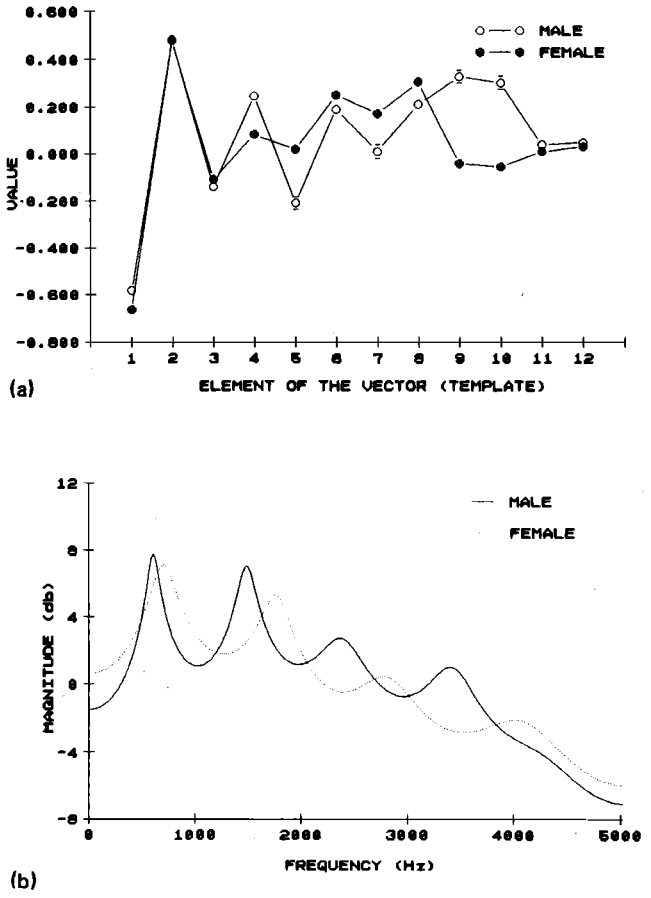

FIG. 3. (a) Two reflection coefficient templates from the gender level in Fig. 2 for male and female speakers. (b) The corresponding spectra.

we obtained averaged templates for each phoneme group (e.g., vowels, unvoiced fricatives, or voiced fricatives) for each speaker for each of the four feature parameter vectors. At the gender level each gender was represented by a single token obtained by averaging all templates from the speaker level for each speaker for a given gender. Thus at the gender level we obtained a phoneme group template for vowels, unvoiced fricatives, and voiced fricatives for each gender for each of the four feature parameter vectors. The amount of template averaging increases from the token level to the gender level in Fig. 2.

Figure 3(a) shows two reflection coefficient (RC) templates for vowels for the gender level in Fig. 2 for male and female speakers. The LPC filter order was 12, giving 12 elements for each template. Each template might be considered a "universal token" representing each gender. The data depicted in Fig. 3(a) indicate the mean $\pm$ standard error (SE), which were calculated from the speaker level templates. Note that in some instances the SE is smaller than the dot (circle) used to indicate the mean value, and thus is not apparent in the figure. Later we will show that when these two tokens were used as reference templates for one of our recognition schemes with a Euclidean distance measure, a 100% correct gender recognition rate was achieved. This result is not unexpected since the data in Fig. 3(a) indicate that the within-gender variation for these reflection coefficients, as represented by SE, was small compared to the between-gender variation. Also note that elements 1, 4, 5, 6, 7, 8, 9, and 10 of these reference templates account for the most between-gender variation, while elements 2, 3, 11, and 12 account for little between-gender variation and thus could be discarded to reduce the dimensionality of the vector. Figure 3(b) shows the spectra corresponding to these two "universal" reflection coefficient templates.

Similarly, Fig. 4(a) and (b) show two reflection coefficient templates from the gender level in Fig. 2 (with the same filter order of 12) and the corresponding spectra for unvoiced fricatives for male and female speakers. As for Fig. 3, the SE is smaller than the dot used to indicate the mean value and thus is not visible in the figure. Later we show that when these two tokens were used as reference templates with the same recognition scheme with a Euclidean distance measure, an 80.8% correct gender recognition rate was achieved. Similarly, Fig. 5(a) and (b) consists of two cepstral coefficient templates (with the same filter order of 12) and the corresponding spectra for voiced fricatives for male and female speakers. Later we show that when these two tokens were used as reference templates with the same recognition scheme as above with a Euclidean distance measure, a 98.1% correct gender recognition rate was achieved. The gender spectra in Figs. 3(b), 4(b), and 5(b) are indicative of basic properties of vowels, unvoiced fricatives, and voiced fricatives, e.g., the energy for vowels is concentrated in the

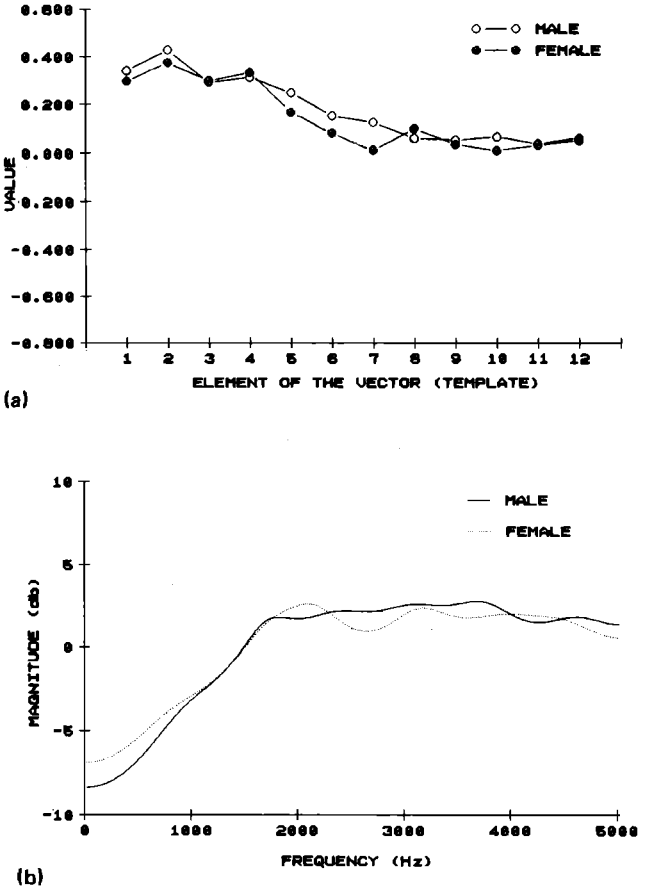

FIG. 4. (a) Two reflection coefficient templates for unvoiced fricatives from the gender level in Fig. 2 for male and female speakers. (b) The corresponding spectra.

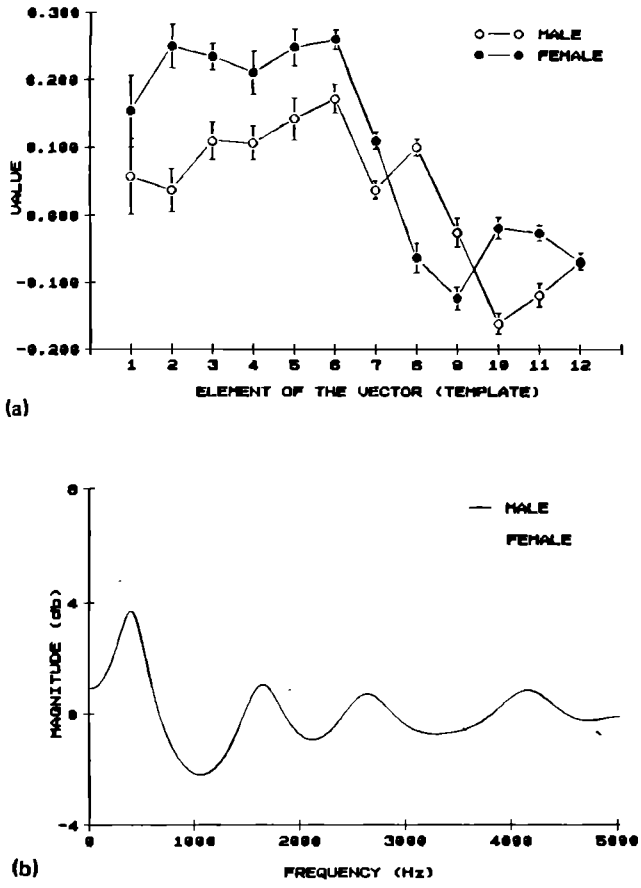

FIG. 5. (a) Two cepstral coefficient templates for voiced fricatives from the gender level in Fig. 2 for male and female speakers. (b) The corresponding spectra.

low-frequency portion of the spectrum, the energy for unvoiced fricatives is concentrated in the high-frequency portion of the spectrum, and the energy for voiced fricatives is more or less equally distributed.

C. Recognition schemes

To investigate the influence of the test and reference template averaging and to establish an objective procedure for gender recognition, several recognition schemes were designed, as summarized in Table I.

Scheme TS (Token template versus Speaker template) is illustrated in Fig. 6(a). In the training stage, one test template for each test utterance (the token level) and one reference template for each subject (the speaker level) were formed. The speaker level set constituted the reference cluster that included all speaker templates. During the testing stage, the distance measure between each token level template for all test speakers and the speaker level templates for each speaker was calculated to determine the minimum distance. The speaker's gender for the token level utterance was classified as male or female according to the speaker level reference template that gave the smallest distance measure. The gender was determined from the speaker's gender.

TABLE I. Four recognition schemes.

	Test template from	Reference template from
Scheme TS	Token level	Speaker level
Scheme TG	Token level	Gender level
Scheme SG	Speaker level	Gender level
Scheme SS	Speaker level	Speaker level

Scheme TG is illustrated in Fig. 6(b). In the training stage, one test template for each test utterance (the token level) and one reference template for each gender (the gender level) were formed. The gender level constituted the reference cluster that included only two gender templates. In the testing stage, the distance measure between each token level template for all test speakers and the two gender level templates was calculated to determine the minimum distance. The speaker's gender for each token level utterance was classified as male or female according to the gender known for the gender level reference template.

Figure 6(c) shows scheme SG. In the training stage, one test template for each test subject (the speaker level) and one reference template for each gender (the gender level) were formed. The speaker level set constituted the test pool that included all speaker templates. In the testing stage the distance measure between each speaker level template for all test speakers and the gender level templates was calculated to determine the minimum distance. The speaker's gender for the speaker level template was classified as male or female according to the gender known for the gender level reference template.

Scheme SS appears in Fig. 6(d). In the training stage, only the speaker level templates were formed with each speaker being represented by a single template. In the testing stage, the leave-one-out or jacknife procedure was applied (see the next section). The distance measure between each speaker level template and the other speaker level templates was calculated to determine the minimum distance. The speaker's gender for the test template was classified as male or female according to the gender known for the reference template. These steps were repeated until all subjects were tested.

In order to use the database effectively, the leave-one-out procedure (Lachenbruch, 1975; Childers, 1989) was adopted for the experiments. For comparison, the resubstitution procedure was also used in selected experiments.

III. EXPERIMENTAL RESULTS

We present the results for all recognition schemes in both tabular and graphical form for both the LPC log likelihood (LLD) measure (Table II and Fig. 7) and the cepstral distance (CD) measure (Table III and Fig. 8). The results for only scheme SG are given for both the Euclidean (EUC) distance measure (Table IV and Fig. 9) and the probability density function (pdf) distance measure (Table V and Fig. 10). In these latter two tables the FFF recognition scores were available only for the LPC filter order of 12. Scheme SG gave the best results for EUC and pdf and these tables and figures provide a more detailed performance evaluation for the various feature vectors. All results in these tables and

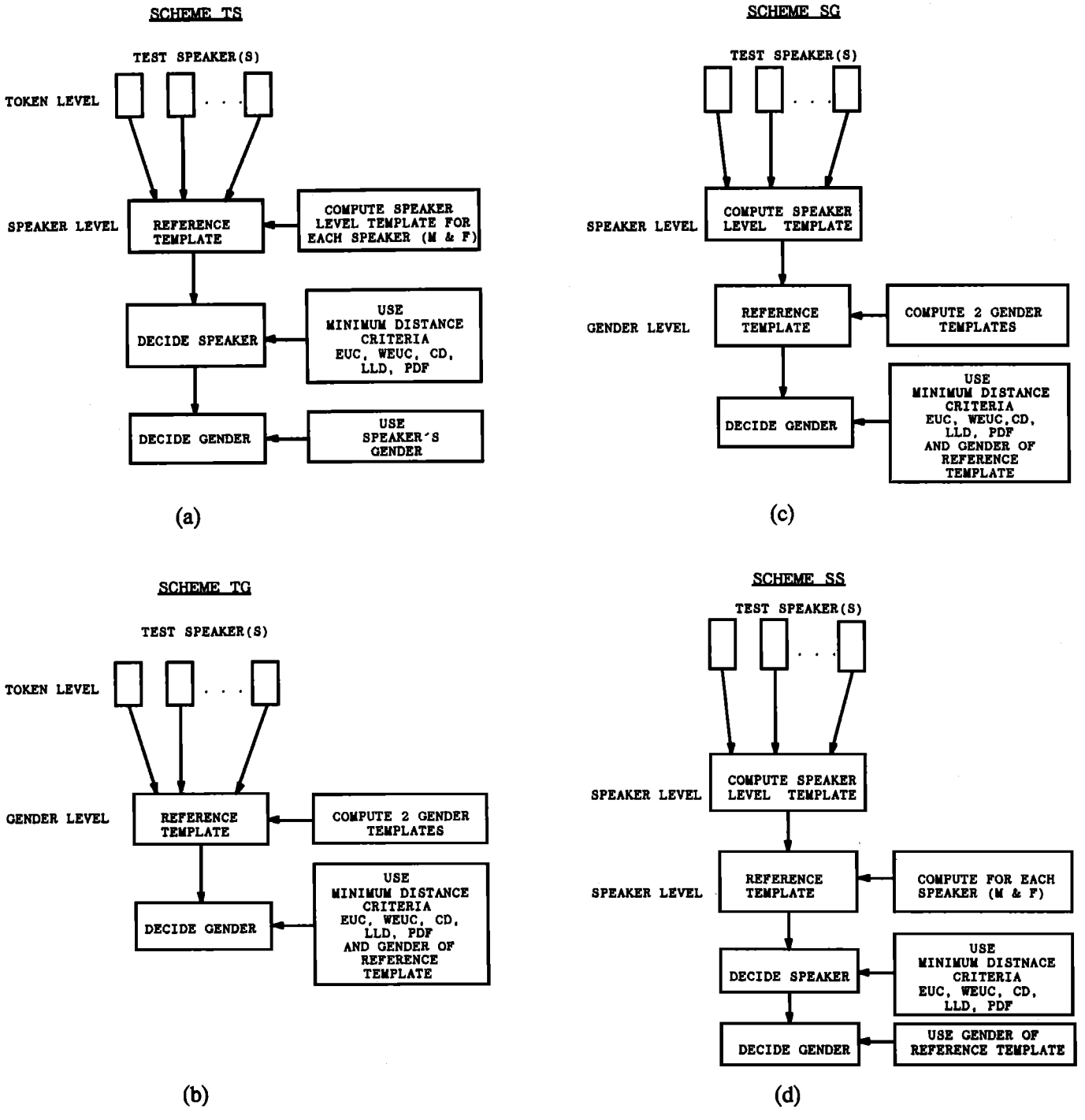

FIG. 6. Four recognition schemes.

figures are for the leave-one-out method and for various filter orders.

Our results showed that most of the LPC-derived feature parameters performed well for gender recognition. Among them, the reflection coefficient (RC) combined with the Euclidean (EUC) distance measure was the best choice for sustained vowels (100%). The cepstral distance measure was the best for unvoiced fricatives, while the LPC log likelihood distance measure, the reflection coefficient combined with the Euclidean distance, and the cepstral distance measure were nearly equal in performance for voiced fricatives. The EUC achieved better results than the pdf. The averaging techniques were important for designing appropriate test and reference templates and a filter order of 12–16 was sufficient for most designs. Since the weighted Euclidean (WEUC) distance measure is a simplified case of the pdf and the results produced by the WEUC were nearly the same as those produced by the pdf in our experiments, only the pdf results will be presented and discussed.

Figures 7 and 8 indicate that the highest recognition

TABLE II. Results for the LPC log likelihood distance measure for various recognition schemes and various filter orders.

		Correct gender recognition rate in %			
		Order = 8	Order = 12	Order = 16	Order = 20
Sustained vowels	Scheme TS	63.1	69.6	74.2	74.2
	Scheme TG	65.2	71.5	76.2	76.5
	Scheme SG	75.0	86.5	86.5	84.6
	Scheme SS	75.0	80.8	86.5	88.5
Unvoiced fricatives	Scheme TS	59.2	64.2	67.7	65.0
	Scheme TG	61.5	63.9	64.2	64.2
	Scheme SG	67.3	75.0	75.0	78.9
	Scheme SS	76.9	75.0	73.1	69.3
Voiced fricatives	Scheme TS	74.5	72.1	73.1	72.6
	Scheme TG	77.4	80.3	81.7	80.3
	Scheme SG	90.4	94.2	96.2	98.1
	Scheme SS	98.1	96.2	96.2	94.3

rates in all cases were achieved using schemes SG and SS. This was true for all filter orders regardless of the type of utterance used. For voiced fricatives the highest correct recognition rate, 98.1%, was obtained with scheme SS using the LPC log likelihood measure with a filter order of 8. The cepstrum coefficients combined with the Euclidean distance measure gave the highest correct recognition rate of 90.4% for unvoiced fricatives with a filter order of 16. The most effective combinations of feature vectors and distance measures for schemes SG and SS are summarized in Table VI.

IV. DISCUSSION

A. Comparison of recognition schemes

The type of template forming and recognition scheme used was important for a high correct gender recognition rate. Template averaging appears critical since the highest recognition rates were obtained using schemes SG and SS, wherein the test and reference templates were formed by averaging all the utterances from the same speaker or the same gender.

TABLE III. Results for the cepstral distance measure for various recognition schemes and various filter orders.

		Correct recognition rate in %			
		Order = 8	Order = 12	Order = 16	Order = 20
Sustained vowels	Scheme TS	61.3	68.3	69.4	70.6
	Scheme TG	69.4	67.3	70.0	72.1
	Scheme SG	82.7	92.3	90.4	90.4
	Scheme SS	90.4	92.3	92.3	88.5
Unvoiced fricatives	Scheme TS	61.2	65.8	63.9	64.6
	Scheme TG	58.8	61.5	62.7	64.2
	Scheme SG	71.2	75.0	84.6	82.7
	Scheme SS	78.8	88.5	90.4	88.5
Voiced fricatives	Scheme TS	79.3	82.7	81.3	80.8
	Scheme TG	75.5	82.2	84.6	85.1
	Scheme SG	94.2	98.1	98.1	96.2
	Scheme SS	92.3	92.3	92.3	90.4

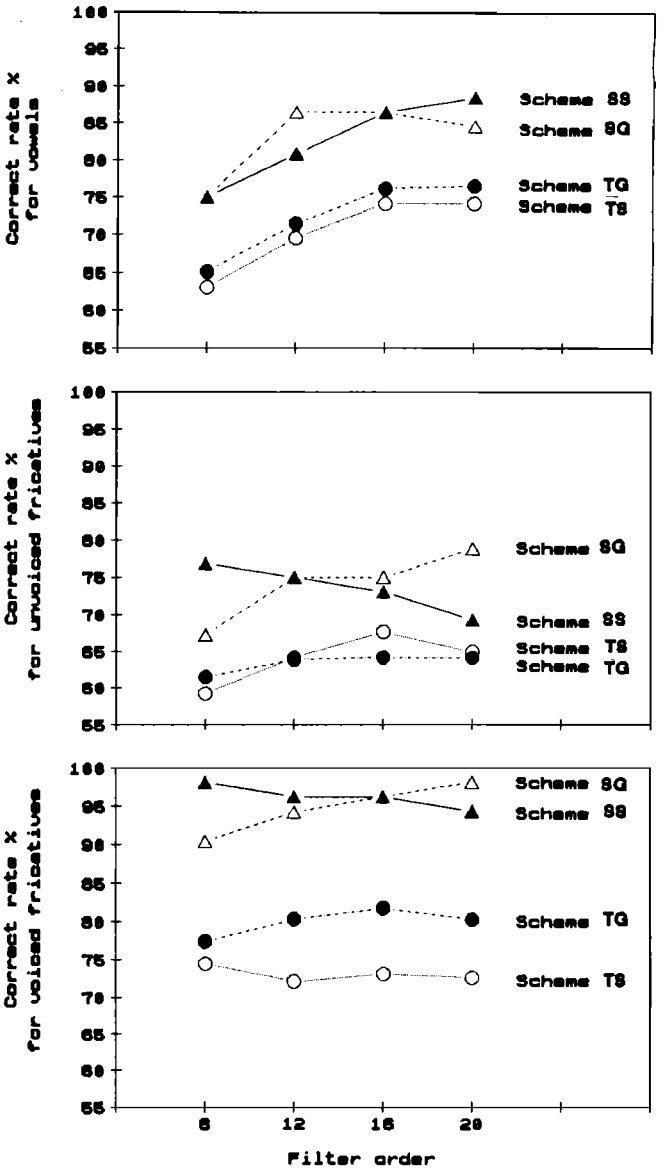

FIG. 7. Results for the LPC log likelihood distance measure for various recognition schemes and various filter orders.

The acoustic feature vectors were extracted from different utterances and different speakers at different times. Consequently, we might reasonably deduce that the gender information we extracted is time invariant, phoneme independent, and speaker independent for a given gender. Averaging appeared to emphasize the speaker's gender information and increase the between-to-within gender variation ratio. This conclusion is consistent with the findings of Lass and Mertz (1978), who decided that temporal cues appeared not to play a role in speaker gender identification. Furthermore, the use of a long-term averaging technique appeared useful for text-independent speaker recognition (Pruzansky, 1963; Markel *et al.*, 1977), which can be argued to include aspects of gender recognition.

Our results indicate that scheme SG and scheme SS are nearly the same in performance. However, from a practical point of view, scheme SG would be easier to implement in an automated system since only two reference templates are needed.

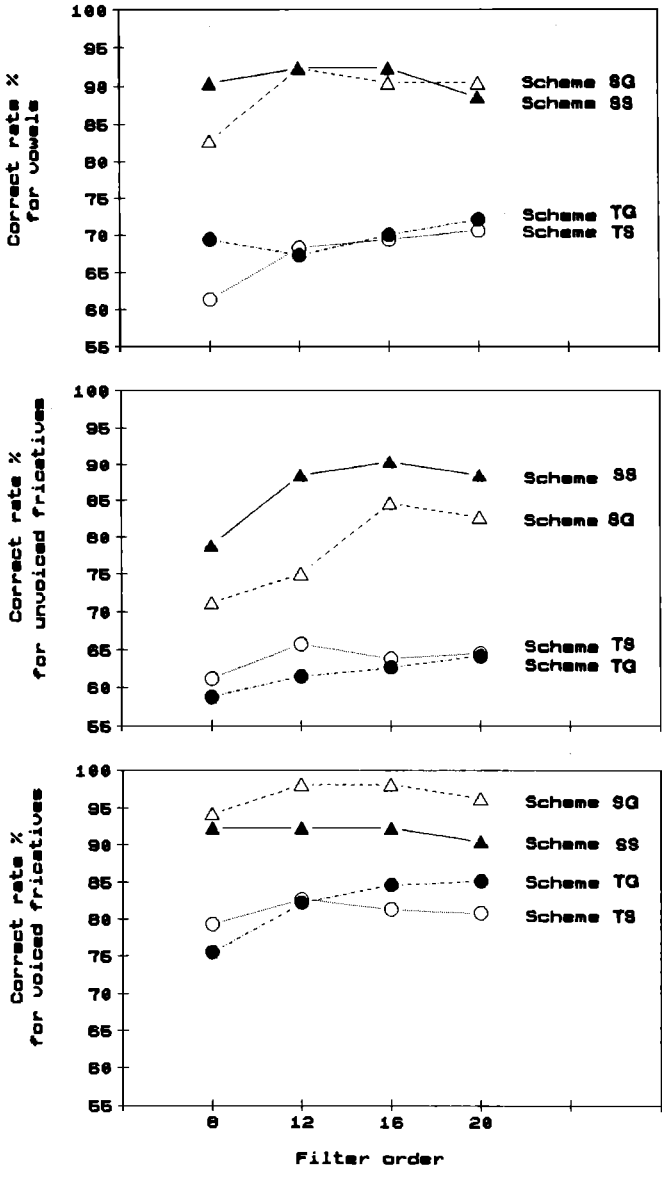

FIG. 8. Results for the cepstral distance measure for various recognition schemes and various filter orders.

B. Comparison of acoustic features: LPC versus cepstrum

Although both LPC log likelihood and cepstral distance measures were effective tools in classifying male/female voices, the performance of the latter was better than the former. Figures 7 and 8 and Table VI show that the cepstrum coefficient features are superior to the LPC coefficients for gender discrimination, with the exception of voiced fricatives, where both sets of coefficients achieved a high of 98.1% correct recognition. The cepstral distance measure performed nearly equally well for both male and female groups, indicating that this measure has some "normalizing characteristics," while the LPC recognizer did not have these characteristics.

TABLE IV. Results for the Euclidean distance measure for recognition scheme SG for various filter orders and the four acoustic feature vectors.

		Correct recognition rate in % Order = 8	Order = 12	Order = 16	Order = 20
Sustained vowels	ARC	78.8	78.8	78.8	82.7
	LPC	73.1	78.8	80.8	80.8
	FFF	N/A	98.1	N/A	N/A
	RC	88.5	100.0	100.0	100.0
	CC	82.7	92.3	90.4	90.4
Unvoiced fricatives	ARC	75.0	75.0	75.0	75.0
	LPC	80.8	69.2	71.2	71.2
	RC	80.8	80.8	80.8	80.8
	CC	71.2	75.0	84.6	82.7
Voiced fricatives	ARC	86.5	88.5	86.5	88.5
	LPC	92.3	92.3	92.3	90.4
	RC	94.2	96.2	96.2	96.2
	CC	94.2	98.1	98.1	96.2

C. Other acoustic parameters

The results of Figs. 9 and 10 and Table VI indicate that the overall performance using RC or CC was better than that achieved using ARC and LPC coefficients, when the EUC distance measure was used. The RC functioned well with sustained vowels, achieving a correct recognition rate of 100% for filter orders of 12, 16, and 20. The CC performed well with unvoiced fricatives, with the highest correct recognition rate being 84.6% with a filter order of 16. Both RC and CC worked well with voiced fricatives, regardless of the filter order. The LPC combined with pdf to achieve correct recognition rates of 98.1%, 86.5%, and 94.2%, for vowels, unvoiced fricatives, and voiced fricatives, respectively, with a filter order of 12. However, this scheme is sensitive to filter order. The results for the fundamental frequency and formant (FFF) parameters are discussed in detail in paper II.

D. Comparison of vowels, unvoiced fricatives, and voiced fricatives

The results of Figs. 7–10 and Table VI indicate that either vowels (100%, scheme SG, vector RC, distance measure EUC) or unvoiced fricatives (90.4%, scheme SS, vector CC, distance measure EUC) or voiced fricatives (98.1%, scheme SG, vector CC, distance measure EUC; 98.1%, scheme SS, vector LPC, distance measure LLD) could be used to objectively classify speaker's gender.

As mentioned earlier, the acoustic parameters used in this coarse analysis were derived using the LPC all-pole model that attempts to match the spectra of the data to that of the model. The LPC log likelihood and cepstral distance measures are directly related to differences of the power spectra of the test and reference signals. Consequently, the results indicate that the spectral characteristics were major factors for distinguishing the speaker's gender. Furthermore, gender recognition was achievable using unvoiced fricatives, indicating that the speaker's gender could be determined from knowledge of the speaker's vocal tract

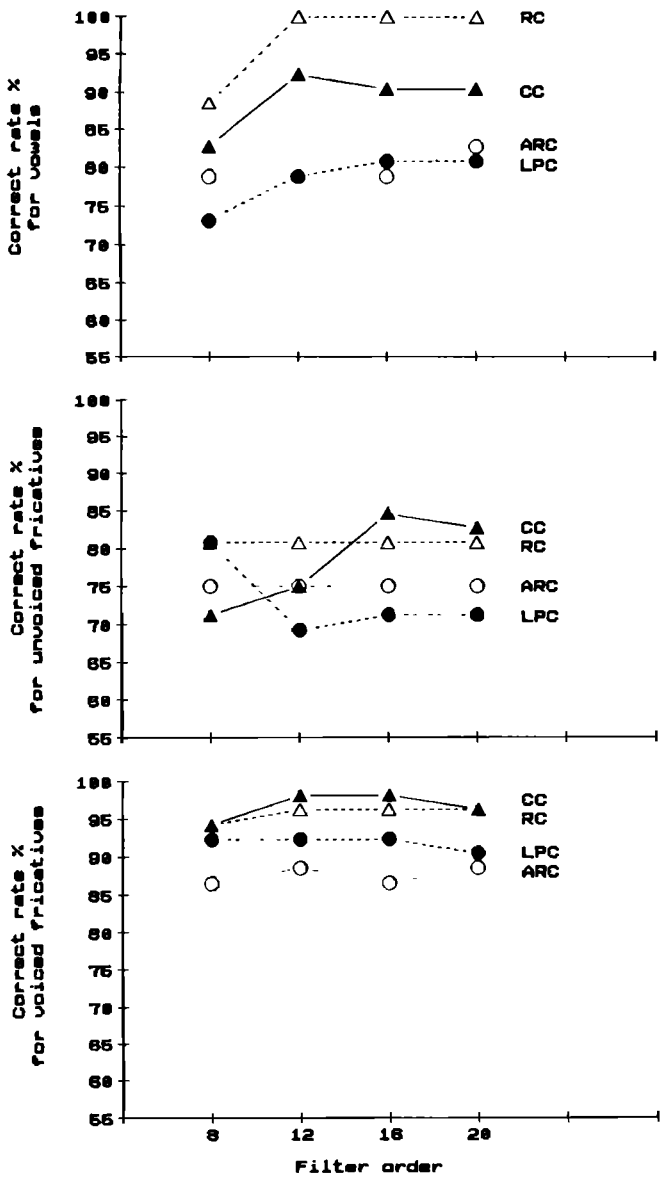

FIG. 9. Results for the Euclidean distance measure for recognition scheme SG for various filter orders and four acoustic feature vectors.

characteristics only, since no voicing information is available for unvoiced fricatives. Moreover, if voicing information was combined with the vocal tract characteristics, as for vowels and voiced fricatives, gender recognition was improved, despite the fact that fundamental frequency information was not used.

E. Filter order

Figure 9 indicates that for the Euclidean distance measure the filter order is not particularly sensitive and may range from 12–16. However, Fig. 10 is a different case, indicating that for pdf the overall trend for vowels was that recognition rates increased from a filter order of 8 to a peak at 12 and then decreased, with the exception of ARC, where the recognition rate reached its peak with a filter order of 16. All acoustic feature vectors (except for LPC) for voiced and unvoiced fricatives showed decreasing correct recognition rates as the filter order increased from 8–20.

TABLE V. Results for the probability density function distance measure for recognition scheme SG for various filter orders and four acoustic feature vectors.

		Correct recognition rate in %			
		Order = 8	Order = 12	Order = 16	Order = 20
Sustained vowels	ARC	80.8	84.6	88.5	67.3
	LPC	84.6	98.1	92.3	80.8
	FFF	N/A	96.2	N/A	N/A
	RC	88.5	98.1	92.3	67.3
	CC	78.8	94.2	90.3	75.0
Unvoiced fricatives	ARC	69.2	65.4	57.7	N/A
	LPC	78.8	86.5	78.8	53.8
	RC	78.8	73.1	67.3	55.8
	CC	80.8	73.1	69.2	57.7
Voiced fricatives	ARC	88.5	86.5	82.7	59.6
	LPC	92.3	94.2	94.2	71.2
	RC	92.3	90.4	90.4	75.0
	CC	92.3	92.3	80.8	71.2

F. Comparison of distance measures

The EUC distance measure has been generally considered as inferior to the pdf because EUC does not normalize the feature vector dimensions with the dimension with the largest value becoming the most significant. In contrast, the pdf approach does normalize the feature vectors with the covariance matrix. However, the pdf approach did not work well in our experiments with the EUC outperforming the pdf, as indicated in Tables IV–VI and Figs. 9 and 10.

A possible explanation for this result may be that the ratio of the available number of subjects per gender to the number of elements (measurements) per feature vector was small. The assumption with the pdf distance measure is that the data are normally (Gaussianly) distributed. Foley (1972) and Childers (1986) pointed out that if the ratio of the available number of samples per class (in this study, number of subjects per gender) to the number of samples per data record (in this study, number of elements per feature vector) is small, then data classification for both training and test sets may be unreliable. This ratio should be on the order of three or larger (Foley, 1972). In our study, the ratios were 3.25 (26/8), 2.17 (26/12), 1.63 (26/16), and 1.3 (26/20) for filter orders of 8, 12, 16, and 20, respectively. Only the value of 3.25 satisfied the above specified requirement. With the exception of the results for a filter order of 8, where the performances of the pdf and EUC were comparable, the pdf approach did not function well. The smaller the ratio, the poorer the pdf results.

G. Variability of female voices

Generally, the performance of the various feature vectors combined with the various distance measures for male subjects was better than that for the female subjects. The statistical analyses of these data indicated that the correct recognition rates for males had larger mean values, larger

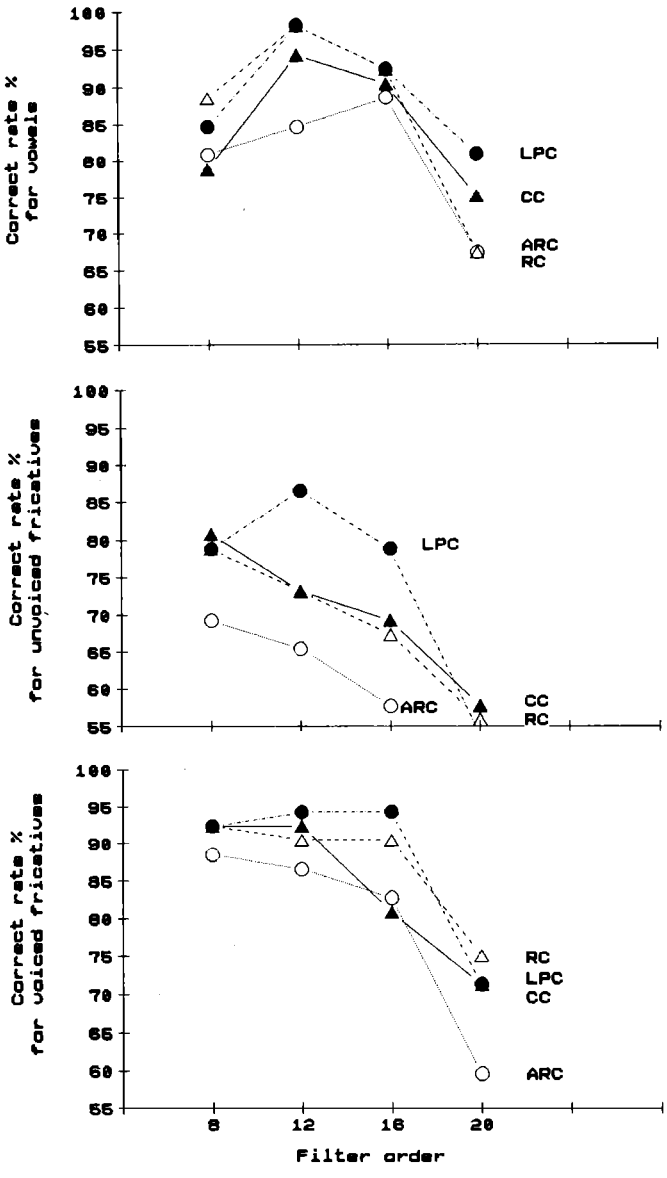

FIG. 10. Results for the probability density function distance measure for recognition scheme SG for various filter orders and four acoustic feature vectors.

minimum values, and smaller standard deviations than those for females, suggesting that female feature vectors had larger variations than the male feature vectors.

To further confirm this conclusion the Wilcoxon signed-ranks test and the paired samples t test (Ott, 1984) were also performed (Wu, 1990). Both tests indicated that there do exist statistically significant differences between the male and female correct recognition rates.

H. Fisher's discriminant ratio criterion

We also analyzed the acoustic feature vectors using Fisher's discriminant ratio to determine the ability of a feature to separate genders using the distance between genders and the scatter within genders (Childers *et al.*, 1982; Parsons, 1986; Wu, 1990). The separation was estimated by representing each gender (class) by its mean and the variance of the means. The variance was then compared to the

TABLE VI. Summary of the most effective feature vector and distance measure combinations for scheme SG (except as noted).

		Correct recognition rate in % Filter order			
		8	12	16	20
Sustained vowels	LPC (pdf)		98.1		
	RC (pdf)		98.1		
	FFF (EUC)		98.1		
	RC (EUC)		100.0	100.0	100.0
Unvoiced fricatives	LPC (pdf)		86.5		
	CC (EUC)[a]		88.5	90.4	88.5
Voiced fricatives	LPC (LLD)			96.2	98.1
	LPC (LLD)[a]	98.1	96.2	96.2	
	CC (EUC)		98.1	98.1	96.2
	RC (EUC)		96.2	96.2	96.2

[a] By using recognition scheme SS.

average width of the distribution for each class (i.e., the mean of the individual variances). This measure is commonly called the F ratio:

$$F = \frac{\text{Variance of the means(over all classes)}}{\text{Mean of the variances(within classes)}}. \quad (2)$$

The F ratio reduces to Fisher's discriminant when it is used for evaluating a single feature and there are only two classes.

In addition we estimated the expected probabilities of errors for the acoustic features ARC, LPC, FFF, RC, and CC for the male and female groups (i.e., classes) for vowels, unvoiced fricatives, and voiced fricatives.

In summary, for a filter order of 12, the analytical inferences from Fisher's discriminant and the expected probabilities of error for the various acoustic feature vectors were comparable to the empirical results of the experiments with the pdf distance measure (Wu, 1990). The implication of this result is that an analytical study using a discriminant function and expected probabilities of error to predict the performance of a feature vector for gender recognition generally agreed with experimental error rates calculated using the same feature vector.

V. CONCLUSIONS

Considering that only approximately 150 ms of the speech signal was used for the experiments, automatic gender recognition appears feasible.

Most of the LPC-derived feature vectors functioned well for gender recognition, with the RC vector combined with the EUC distance measure providing 100% correct recognition rate for sustained vowels. The CD measure for unvoiced fricatives achieved 90.4% correct gender recognition while several methods gave 98.1% correct gender recognition for voiced fricatives. The most effective feature vectors and distance measures are summarized in Table VI.

Spectral characteristics were important factors for recognizing the speaker's gender. Either vowels, unvoiced fricatives, or voiced fricatives could be used to classify the subject's gender objectively. The speaker's gender features were

represented by only the speaker's vocal tract characteristics when unvoiced fricative data were used. When both source and tract characteristics were represented in the feature vectors (vowels or voiced fricatives) then gender discrimination improved.

Template forming techniques and recognition schemes were important for achieving high correct recognition rates. Recognition schemes SG and SS were better for gender discrimination than schemes TS and TG, indicating the importance of averaging techniques. In addition, averaging techniques seemed more useful than clustering techniques. To a great extent, averaging both test and reference templates eliminated the intrasubject variation within different vowels or fricatives for a given speaker and emphasized features representing the speaker's gender. We conclude that on the whole gender information appears to be time invariant, phoneme independent, and speaker independent for a given gender.

The performance of the CD measure was better than that of the LLD distance measure. Further, the CD measure performed equally well for males and females, indicating that this measure performed some data "normalization." The EUC distance measure was more effective than the pdf. Filter orders of 12–16 were the most appropriate for the majority of design options.

The performance of various feature vectors combined with various distance measures for male subjects were generally better than for female subjects, indicating that female feature vectors had a higher variability than male feature vectors.

An analytical study of the feature vectors using discriminant analysis and expected probabilities of error provided comparable results to our experimental findings.

Improvements in our results can undoubtedly be obtained, e.g., the number of data frames used to calculate templates can probably be reduced, especially for sustained vowels. The feature vector dimensionality can be reduced by eliminating those elements that account for little between-class variation. For example, Fig. 3(a) shows two "universal" reflection coefficient templates (gender level) for male and female speakers. Elements 2, 3, 11, and 12 of these reference templates accounted for little between-gender variation. Consequently, these elements could be discarded to reduce the dimensionality of the vector. Instead of using equal weighting factors in the averaging operation, different weighing factors could be applied to different phoneme feature vectors according to the probability of the phoneme appearance in a real speech situation. This would probably better approximate time averaging than was done in this study. A k th-order nearest neighbor (KNN) decision procedure could be developed to further improve the gender recognition rate.

ACKNOWLEDGMENTS

This research was supported in part by NSF Grant No. ECE-8413583 and NIH Grant No. NIDCD DC 00577 with additional support from the University of Florida Center of Excellence Program in Information Transfer and Processing and the Mind–Machine Interaction Research Center.

Atal, B. S. (1974). "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," J. Acoust. Soc. Am. **55**, 1304–1312.

Atal, B. S. (1976). "Automatic recognition of speakers from their voices," Proc. IEEE **64**, 460–475.

Carlson, T. E. (1981). "Some acoustical and perceptual correlates of speaker gender identification," Ph.D. dissertation proposal, University of Florida, Gainesville.

Cheng, Y. M., and Guerin, B. (1987). "Control parameters in male and female glottal sources," in *Laryngeal Function in Phonation and Respiration*, edited by T. Bear, C. Sasaki, and K. Harris, (College Hill, San Diego, CA) Chap. 17, pp. 219–238.

Childers, D. G. (1986). "Single-trial event-related potentials: statistical classification and topography," in *Topographic Mapping of Brain Electrical Activity*, edited by F. H. Duffy (Butterworths, Boston, MA), Chap. 14, pp. 255–277.

Childers, D. G. (1989). "Biomedical signal processing," in *Selected Topics in Signal Processing*, edited by S. Haykin (Prentice-Hall, Englewood Cliffs, NJ), Chap. 10, pp. 194–250.

Childers, D. G., Bloom, P. A., Arroyo, A. A., Roucos, S. E., Fischler, I. S., Achariyapaopan, T., and Perry, N. W., Jr. (1982). "Classification of cortical responses using features from single EEG records," IEEE Trans. Biomed. Eng. **BME-29**, 423–438.

Childers, D. G., and Wu, K. (1990). "Quality of speech produced by analysis–synthesis," Speech Commun. **9**, 97–117.

Childers, D. G., Wu, K., and Hicks, D. M. (1987). "Factors in voice quality: Acoustic features related to gender," Proc. IEEE Int. Conf. Acoust. Speech Signal Process. **1**, 293–296.

Childers, D. G., Wu, K., Bae, K. S., and Hicks, D. M. (1988). "Automatic gender recognition of gender by voice," Proc. IEEE Int. Conf. Acoust. Speech Signal Process. **1**, 603–606.

Childers, D. G., and Wu, K. (1991). "Gender recognition from speech. Part II: Fine analysis," J. Acoust. Soc. Am. **90**, 1841–1856.

Coleman, R. O. (1976). "A comparison of the contributions of two voice quality characteristics to the perception of maleness and femaleness in the voice," J. Speech Hearing Res. **19**, 168–180.

Committee on Evaluation of Sound Spectrograms. (1979). "On the theory and practice of voice identification," National Academy of Sciences Report, Washington, DC.

Davis, S., and Mermerlstein, P. (1980). "Comparison of parametric representation for monosyllable word recognition in continuously spoken sentences," IEEE Trans. Acoust. Speech and Signal Process. **28**, 375–366.

Fant, G. (1976). "Vocal tract energy functions and non-uniform scaling," J. Acoust. Soc. Jpn. **11**, 1–18.

Fant, G. (1979). "Glottal source and excitation analysis." Speech Trans. Lab.-Q. Prog. Status Rep. **4**, 85–107.

Foley, D. H. (1972). "Considerations of sample and feature size," IEEE Trans. Inform. Theor. **IT-18**, 618–626.

Furui, S. (1989). *Digital Speech Processing, Synthesis, and Recognition* (Dekker, New York).

Gray, A. H., and Markel, J. D. (1976). "Distance measures for speech processing," IEEE Trans. Acoust. Speech Signal Proc. **24**, 380–391.

Gray, R., Buzo, A. H., and Matusyama, Y. (1980). "Distortion measures for speech processing," IEEE Trans. Acoust. Speech and Signal Process. **24**, 380–391.

Henton, C. G. (1987). "Fact and fiction in the description of female and male pitch," J. Acoust. Soc. Am. Suppl. 1 **82**, S91.

Hirano, M., Kurita, J., and Nakahima, T. (1983). "Growth, development, and aging of human vocal folds," in *Vocal Fold Physiology: Contemporary Reseach and Clinical Issues*, edited by D. Bless and J. Abbs (College-Hill, Press, San Diego, CA), pp. 22–43.

Hollien, H., and Jackson, B. (1973). "Normative data on the speaking fundamental frequency characteristics of young adult males," J. Phon. **1**, 117–120.

Hollien, H., and Malcik, E. (1967). "Evaluation of cross-sectional studies of adolescent voice changes in males," Speech Monogr. **34**, 80–84.

Hollien, H., and Paul, P. (1969). "A second evaluation of the speaking fundamental frequency characteristics of post-adolescent girls," Language Speech **12**, 119–124.

Hollien, H., and Shipp, T. (1972). "Speaking fundamental frequency and chronologic age in males," J. Speech Hear. Res. **15**, 155–159.

Holmberg, E. B., Hillman, R. E., and Perkell, J. S. (1988). "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice," J. Acoust. Soc. Am. **84**, 511 529.

in soft, normal, and loud voice," J. Acoust. Soc. Am. **84**, 511–529.
Horri, Y., and Ryan, W. J. (**1981**). "Fundamental frequency characteristics and perceived age of adult male speakers," Folia Phoniatr. **33**, 227–233.
Itakura, F. (**1975**). "Minimum prediction residual principle applied to speech recognition," IEEE Trans. Acoust. Speech Signal Process. **ASSP-23**, 67–72.
Juang, B. H. (**1984**). "On using the Itakura–Saito measures for speech coder performance evaluation," AT&T Bell Lab. Technol. J. **63**, 1477–1498.
Klatt, D. H. (**1987**). "Acoustic correlates of breathiness: First harmonic amplitude, turbulence noise, and tracheal coupling," J. Acoust. Soc. Am. Suppl. 1 **82**, S91.
Klatt, D. H., and Klatt, L. C. (**1990**). "Analysis, synthesis and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am. **80**, 820–857.
Lachenbruch, P. A. (**1975**). *Discriminaut Analysis* (Hafner, New York).
Lass, N. J., and Mertz, P. J. (**1978**). "The effect of temporal speech alterations on speaker race and sex identifications," Lang. Speech **21**, 279–290.
Linke, C. E. (**1973**). "A study of pitch characteristics of female voices and their relationship to vocal effectiveness," Folia Phoniatr. **25**, 173–185.
Linville, S. E., and Fisher, H. B. (**1985**). "Acoustic characteristics of perceived versus actual vocal age in controlled phonation by adult females," J. Acoust. Soc. Am. **78**, 40–48.
Markel, J. D., Oshika, B., and Gray, A. H., Jr. (**1977**). "Long-term feature averaging for speaker recognition," IEEE Trans. Acoust. Speech Signal Process. **ASSP-25**, 330–337.
Makhoul, J. (**1975**). "Linear prediction in automatic speech recognition," in *Speech Recognition*, Invited papers presented at the 1974 IEEE Symposium, edited by D. R. Reddy (Academic, New York), pp. 183–220.
Monsen, R. B., and Engebretson, A. M. (**1977**). "Study of variations in the male and female glottal wave," J. Acoust. Soc. Am. **62**, 981–993.
Monsen, R. B., and Engebretson, A. M. (**1983**). "The accuracy of formant frequency measurements: a comparison of spectrographic analysis and linear prediction," J. Speech Hear. Res. **26**, 89–97.
Murry, T., and Singh, S. (**1980**). "Multidimensional analysis of male and female voices," J. Acoust. Soc. Am. **68**, 1294–1300.
Nocerino, N., Soong, F. K., Rabiner, L. R., and Klatt, D. H. (**1985**). "Comparative study of several distortion measures for speech recognition," Speech Commun. **4**, 317–331
O'Kane, M. (**1987**). "Recognition of speech and recognition of speaker sex: parallel or concurrent processes?" J. Acoust. Soc. Am. Suppl. 1 **82**, S84.
Ott, L. (**1984**). *An Introduction to Statistical Methods and Data Analysis* (Duxbury, Boston, MA).
Parsons, T. W. (**1986**). *Voice and Speech Processing* (McGraw-Hill, New York).
Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184.
Pruzansky, S. (**1963**). "Pattern matching procedure for automatic talker recognition," J. Acoust. Soc. Am. **35**, 354–358.
Rabiner, L. R., and Levinson, S. E. (**1981**). "Isolated and connected word recognition—theory and selected applications," IEEE Trans. Commun., **Com-29**, 621–659.
Rabiner, L. R., and Schafer, R. W. (**1978**). *Digital Processing of Speech Signal* (Prentice–Hall, Englewood Cliffs, NJ).
Rosenberg, A. E. (**1976**). "Automatic speaker verification: a review," Proc. IEEE **64**, 475–487.
Rothenberg, M. (**1981**). "Acoustic interaction between the glottal source and the vocal tract," in *Vocal Fold Physiology*, edited by K. N. Stevens and M. Hirano (Univ. of Tokyo Press, Tokyo), pp. 305–328.
Saxman, J., and Burk, K. (**1967**). "Speaking fundamental frequency characteristics of middle-aged females," Folia Phoniatr. **19**, 167–172.
Schwartz, M. F. (**1968**). "Identification of speaker sex from isolated, voiceless fricatives," J. Acoust. Soc. Am. **43**, 1178–1179.
Singh, S., and Murry, T. (**1978**). "Multidimensional classification of normal voice qualities," J. Acoust. Soc. Am. Suppl. 1 **64**, S81–S87.
Stoicheff, M. (**1981**). "Speaking fundamental frequency characteristics of non-smoking female adults," J. Speech Hear. Res. **24**, 437–441.
Titze, I. R. (**1987**). "Physiology of the female larynx," J. Acoust. Soc. Am. Suppl. 1 **82**, S90.
Titze, I. R. (**1989**). "Physiologic and acoustic differences between male and female voices," J. Acoust. Soc. Am. **85**, 1699–1707.
Wu, K. (**1990**). "Towards Automatic Gender Recognition from Speech," Ph.D. dissertation, University of Florida.