Principal-components analysis for low-redundancy encoding of speech spectra

Stephen A. Zahorian

Old Dominion University, Norfolk, Virginia 23508

Martin Rothenberg

Syracuse University, Syracuse, New York 13210

(Received 3 April 1980; accepted for publication 5 November 1980)

The principal-components statistical procedure for data reduction is used to efficiently encode speech power spectra by exploiting the correlations of power spectral amplitudes at various frequencies. Although this data-reduction procedure has been used in several previous studies, little attempt was made to optimize the methods for spectral selection and coding through the use of intelligibility testing. In the present study, principal-components basis vectors were computed from the continuous speech of several male and female speakers using various nonlinear spectral amplitude scales. Speech was synthesized using a combination linear predictive (LP) principal-components vocoder. Of the amplitude scales investigated for use with a principal-components analysis of speech spectra, logarithmic amplitude coding of non-normalized spectra emerged as a slight favorite. Speech synthesized from four principal components was found to be about 80% intelligible using a form of the Diagnostic Rhyme Test for rhyming word pairs and about 95% intelligible for words within a sentence context. Speech synthesized from spectral principal components compared favorably in intelligibility and quality with speech synthesized from a control LP vocoder with the same number of parameters.

PACS numbers: 43.70.Ep, 43.70.Gr, 43.70.Lw, 43.70.Jt

INTRODUCTION

The principal-components statistical data-reduction procedure has often been used for efficient encoding of speech spectra. The procedure exploits the experimentally observed correlations among spectral band energies at different frequencies in order to derive a much smaller set of statistically independent parameters (principal components) which retain most of the information present in the original speech spectra. The principal components can be regarded as spectral shape factors which, for a given number of principal components, best explain the overall shape of the spectra. This type of description contrasts with a formant description in which more emphasis is placed on the major spectral peaks and less on the overall spectral shape.

The principal-components statistical procedure has been used previously by several researchers to remove redundancy from speech spectral data. The present study is an extension of this earlier work; however, unlike any previous work in this area, we have also quantitatively evaluated some of the important variables in the procedure through the use of comprehensive speech intelligibility and quality testing. Also unique to the present study is the use of a linear predictive vocoder, rather than a channel vocoder, as the fundamental speech analysis–synthesis tool. The linear prediction vocoder was selected because of its well-defined mathematical model, and also because of its ease of implementation and computational efficiency.

Kramer and Mathews (1956) were apparently the first researchers to utilize the correlations among the various channels of a channel vocoder to obtain a more efficient coding of speech spectra. As a starting point, they used linear amplitude codings of the speech spectral band energies, rather than a more perceptually relevant amplitude coding, such as a logarithmic or power-function coding. The Kramer and Mathews (1956) study was also based on the correlation matrix (which incorporates the data-set mean values) rather than the covariance matrix (which does not include the data-set mean values). Later in this paper, we show that data reduction based on the correlation matrix is inherently somewhat inferior to data reduction based on the covariance matrix. Nevertheless, they reported synthesizing fairly intelligible speech using six to ten independent parameters derived from the correlation matrix. They apparently viewed their procedure as a method for efficient transmission of correlated signals, rather than an attempt to analyze the underlying structure of the data set.

The next study which investigated the correlations of channel vocoder signals in order to achieve an efficient coding of speech spectra was reported by Kulya (1964). This study was similar to that of Kramer and Mathews (1956) in that both a linear amplitude coding of the vocoder signals was used, and the statistical properties of the spectral data were summarized in the correlation matrix. Kulya (1964) reported that the vocoder signals could be represented by eight optimum orthogonal parameters (principal components) to within 7.5% of the original vocoder signals (in terms of normalized mean-square error). Kulya (1964) also investigated a "harmonic" vocoder in which the amplitude spectrum is expanded in a Fourier series expansion and observed that the optimum orthogonal parameter set is only slightly more efficient for representing speech spectra than the harmonic vocoder parameter set (not surprising, since the principal-components basis vectors tend to look like sine and cosine functions).

0001-4966/81/030832-14$00.80

Crowther and Rader (1966) were first to report using transformations of log-coded amplitudes of vocoder band energies to achieve an efficient coding of the band energies. Because of the simplicity of implementation, they used Hadamard transform linear combinations, similar to principal components, but not optimum in the mean-square error sense. They found that speech synthesized from Hadamard-transformed vocoder signals encoded with 1850 bits/s was as clear as speech synthesized from the original vocoder signals at a bit rate of 4000 bits/s.

Boehm and Wright (1968) and Li *et al.* (1969) used statistical methods to both reduce the redundancy of speech spectra and simultaneously obtain an efficient analysis tool for the examination of speech spectra. Boehm and Wright calculated eigenvectors of the correlation matrix of the mel–sone (perceptual units of frequency and amplitude, respectively) encoded speech spectra. Li *et al.* used eigenvectors of the variance–covariance matrix corresponding to high-frequency preemphasized log-coded spectra. Although Boehm and Wright were able to reestimate their original spectral data from a small number of dimensions with an apparently much lower average mean-square error than were Li *et al.*, it is difficult to directly compare these two studies because of the procedural differences and the lack of evaluation criteria more independent of the method than is the mean-square error (as, for example, the quality of synthesized speech).

Much of the recent work in the area of low-dimensionality representations of speech spectra has been done by researchers at the Institute for Perception TNO [Plomp *et al.* (1967); Pols *et al.* (1969, 1973); Klein *et al.* (1970); Pols (1971, 1975, 1977); and Nierop *et al.* (1973)]. In all cases, the statistical properties of speech were summarized using a variance–covariance matrix corresponding to level-normalized, log-coded speech spectra, although there is no firm experimental justification for this particular coding choice. The emphasis of their work has been with vowel spectra (Plomp *et al.*, 1967; Pols *et al.*, 1969; Klein *et al.*, 1970; Pols, 1971; Nierop *et al.*, 1973). They noted that a plot of vowels in the space spanned by the first two dimensions is very similar to a plot of the vowels in the $F1, F2$ plane (Plomp *et al.*, 1967). This group has also done some speech synthesis using various numbers of spectral dimensions and a channel vocoder type synthesizer (Pols, 1975). They have reported intelligibility scores of about 50% to 60% for CVC words using four or five dimensions with their particular method.

Sambur (1975) applied the principal-components method to the log-area ratios of a linear predictive vocoder as a method for efficiently coding these parameters. Log-area ratios can be related in a straightforward manner to the cross-sectional areas of a nonuniform acoustic tube approximation of the vocal tract (Atal and Hanauer, 1971), and therefore might be expected to form a natural characterization of voice information. Sambur (1975) reported synthesizing good quality speech with six orthogonal parameters when the statistics were analyzed separately for each sentence and each speaker. Whether or not this method would have worked well when data from a larger piece of text and a large number of speakers was grouped together is unclear since the perceptual significance of log-area ratios is not nearly as well understood as is the perceptual significance of spectral band energies.

All of these studies support the general idea that a principal-components analysis is a useful method both for efficient coding of spectral data and for use in modeling the underlying structure of the spectral data. However, it is difficult to use data from these studies to compare the various versions of principal-components analysis (type of spectral coding, method of spectral selection, etc.) since the data available are always in the form of an error criterion with respect to the particular measurement scale used. Very little effort has been directed to optimizing the procedure through a measure that is independent of the method, such as perceptual testing of synthesized speech.

The most important underlying assumption in a principal-components analysis of speech is that average mean-square error is a good perceptual distance measure for speech spectra. This assumption, from another viewpoint, is that data variance is equivalent to data "information." However, the validity of the assumption depends strongly on the proper scaling of the data. In our work, we have attempted to optimize a low-redundancy principal-components spectral characterization by measuring speech spectra with a variety of scales selected to maximize the likelihood that low mean-square error would correlate well with high intelligibility.

I. THE STATISTICAL PROCEDURE

A. Principal-components method of data reduction

The principal-components method is a general statistical procedure for finding an efficient representation of a set of correlated data. From a geometric viewpoint, this procedure can be seen as translating and rotating the coordinate system used to measure the data. Alternately, the procedure can be considered as deriving an optimal set of orthonormal basis vectors (Karhunen–Loève) for representing the data. The general principal-components method is discussed in most advanced statistics textbooks (for example, Harman, 1976); the principal-components procedure applied to speech spectral data is discussed in the paper by Li *et al.* (1969).

The essential details of the analysis method can be summarized rather briefly as follows. The statistical properties of the original data set (20 band energies sampled once every 12.8 ms for the present study) are contained in the covariance matrix [C] with each element given by

$$C_{ij}=\frac{1}{K}\left[\sum_{k=1}^{K}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)\right], \quad \text{for } i,j=1,2,\ldots,n,$$

where K is the total number of data frames, x_{ki} is the ith data sample of the kth frame, $\bar{x}_i$ is the average over

k (time) of the ith data sample, and n is the number of data elements in each frame.

The principal-components basis vectors are the m eigenvectors ($m \leq n$) of the covariance matrix corresponding to the m largest eigenvalues of the matrix. Each principal component can be obtained by a weighted average of the components of the original data vector, with weighting coefficients given by the corresponding eigenvector. Furthermore, the original data can be reestimated from linear combinations of the principal components, plus average value terms which depend on the original data-set average values. The average value terms are given by

$$M_j = \sum_{i=m+1}^{n} A_{ij} \sum_{p=1}^{n} A_{ip}\bar{x}_p, \quad j = 1, 2, \ldots, n,$$

where A_{ij} = jth component of the ith eigenvector of [C] with the eigenvectors ranked in order according to decreasing eigenvalues. This data-reduction and reconstitution procedure is illustrated in Fig. 1. The procedure as just outlined insures that, for a given number of principal components, the average mean-square error between the original and reconstituted data is minimized.

Kramer and Mathews (1956), as well as Kulya (1964) and Boehm and Wright (1968), formulated the data-reduction procedure in a somewhat different manner than just described, in that the constant value terms indicated in the right-hand box of Fig. 1 were not allowed. In their formulation, the optimal transformation coefficients are obtained from eigenvectors of the correlation matrix, which corresponds to the covariance matrix used in our study, except the data-set average values are not subtracted in forming the matrix. Due to the somewhat more restrictive problem definition (allowing only a rotation of the coordinate system rather than both a translation of the origin and rotation of the coordinate system), the correlation matrix method will usually give a somewhat larger mean-square error, for a given number of dimensions, than will data reduction based on the covariance matrix. Thus the principal-components data-reduction method is presumably a more efficient data-reduction procedure than is the correlation data-reduction procedure described by Kramer and Mathews (1956).

B. Orthogonal rotation to congruence

The principal-components basis vector set is not unique in that there are an infinite number of orthogonal rotations of the principal-components basis vectors which will span the same space. The principal-components basis vectors are unique in that as much as possible of the original data set is accounted for by the first basis vector, as much as possible of the remaining variance is accounted for by the second basis vector, and so on. However, after it has been decided that a certain number of basis vectors are required, the same total variance can be accounted for with a rotated version of the original vectors. The difference is that variance associated with the individual rotated basis vectors will not be the same as variance associated with individual vectors of the original set.

Thus it is possible that two basis vector sets obtained from separate data sets, as for example different speakers, may not appear to closely resemble one another although they actually span the same space. This will occur when the two sets are linearly dependent, i.e., one basis vector set is a slightly rotated version of the other set. When comparing sets of basis vectors obtained from separate speakers, we always normalized the eigenvectors of each speaker by "orthogonal rotation to congruence." This well-defined mathematical procedure, described in detail by both Schoneman (1966) and Cliff (1966), minimizes superficial differences between basis vector sets. Figures 2 and 3 illustrate an example of the apparent differences in similarity between two basis vector sets before and after rotation to congruence. In Fig. 2, which depicts the basis vector sets of two speakers as determined initially, apparent similarities are small, whereas in Fig. 3, which shows the same basis vectors after rotation to congruence, similarities are obvious.

II. METHODOLOGY

A. Spectral selection and coding

The particular scales used for encoding spectral band energies were (A) non-normalized logarithmic, (B) normalized logarithmic, (C) normalized $\frac{1}{3}$ power function, (D) non-normalized $\frac{1}{3}$ power function, and (E) non-normalized linear.

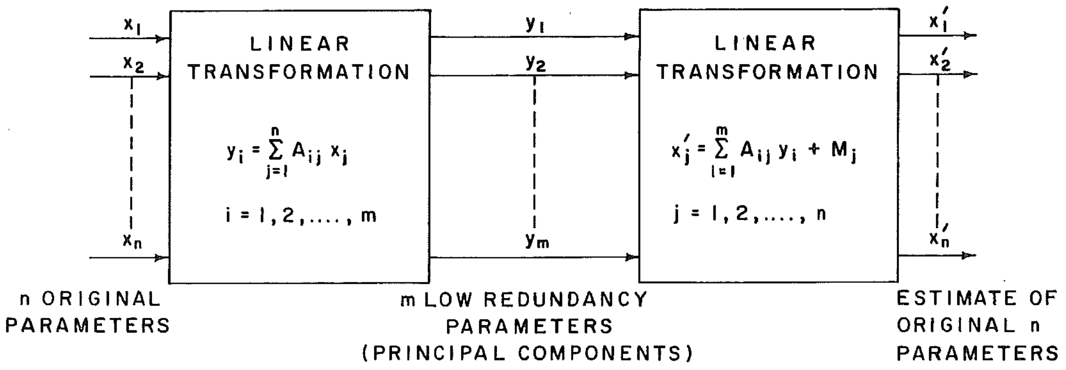
x_1
x_2
x_n
LINEAR TRANSFORMATION
$y_i = \sum_{j=1}^{n} A_{ij} x_j$
$i = 1, 2, \ldots, m$
y_1
y_2
y_m
LINEAR TRANSFORMATION
$x'_j = \sum_{i=1}^{m} A_{ij} y_i + M_j$
$j = 1, 2, \ldots, n$
x'_1
x'_2
x'_n
n ORIGINAL PARAMETERS
m LOW REDUNDANCY PARAMETERS (PRINCIPAL COMPONENTS)
ESTIMATE OF ORIGINAL n PARAMETERS

FIG. 1. Representation of a principal-components low redundancy coding and decoding system. The A_{ij} and M_j are determined from from the statistical properties of the input parameters.

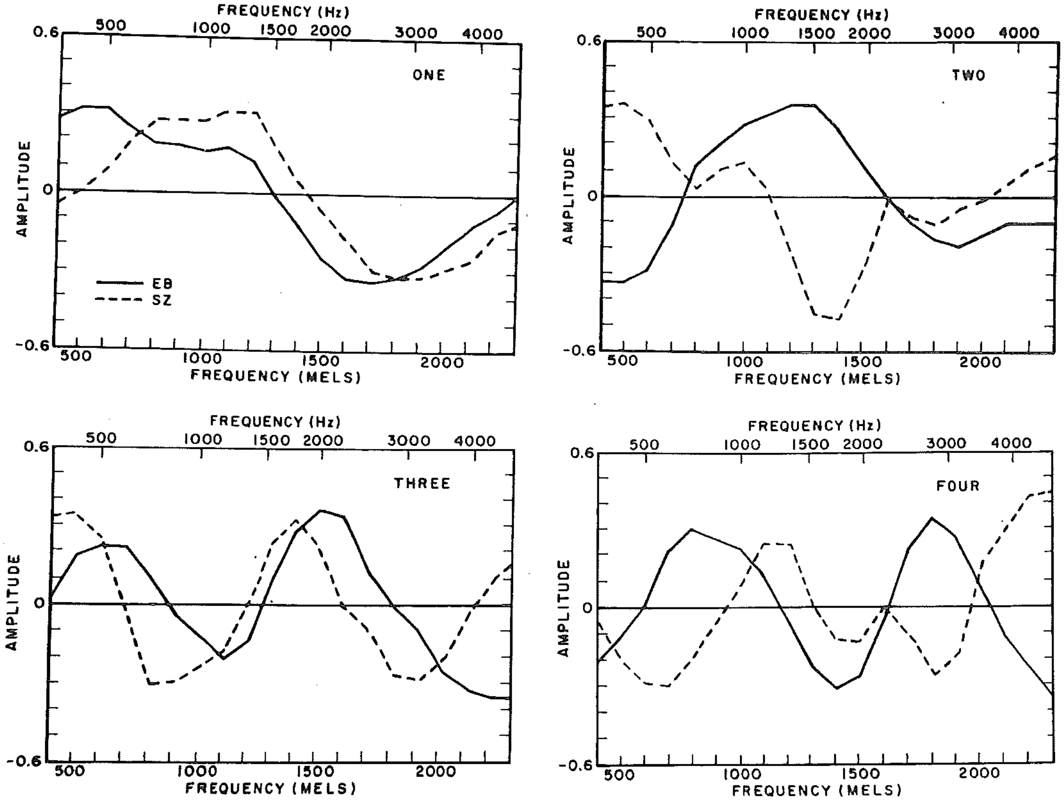

FIG. 2. First four principal-components basis vectors of two speakers before orthogonal rotation to congruence. Basis vectors were obtained from logarithmically amplitude-coded normalized band energies.

These scales were selected on the basis of their relationship to the sone ratio scale of loudness (Stevens, 1936). The $\frac{1}{3}$ power-function scale is a good first approximation to the sone scale (Stevens, 1957), and therefore equal ratios on the power-function scale will correspond to equal loudness ratios. Logarithmic scaling implies that equal distances on the amplitude scale correspond to equal loudness ratios. Logarithmic scaling, commonly used for scaling psychophysical sensations since first proposed by Fechner (1860), also has the property that jnd's (just noticeable differences) in loudness are about equal distances on the scale throughout the range of intensities for most speech sounds. Linear scaling, tested less thoroughly than the other codings, was included in the study primarily to test the effects of a clearly nonperceptually appropriate

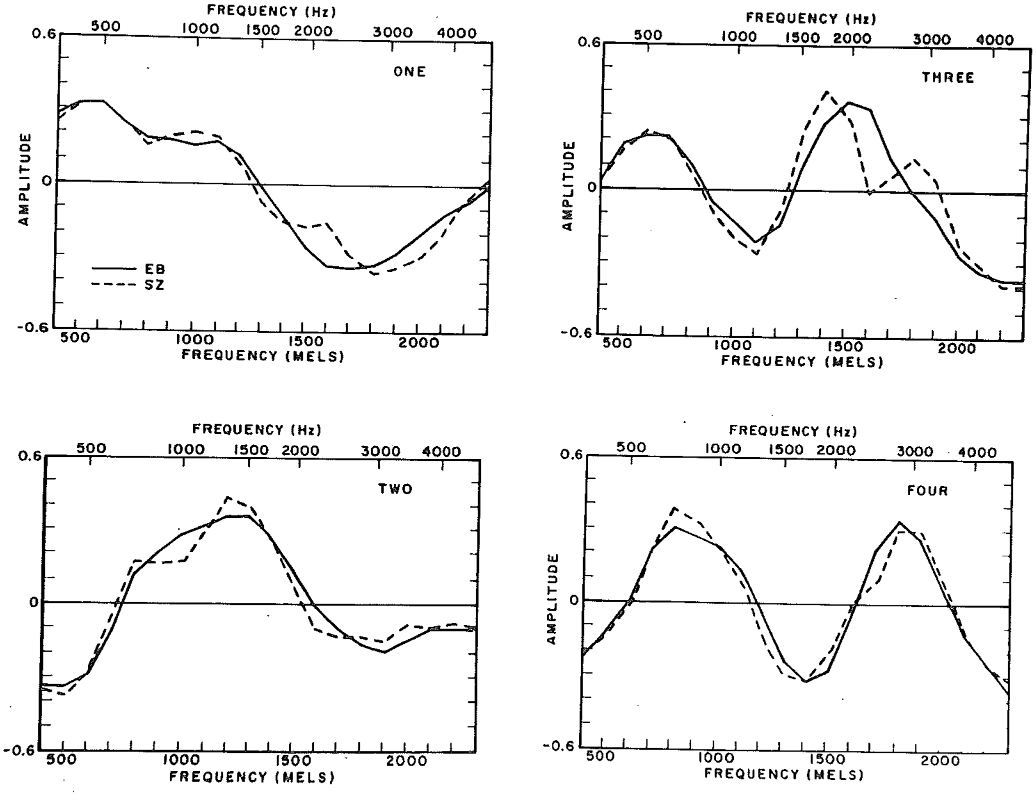

FIG. 3. First four principal-components basis vectors of two speakers after orthogonal rotation to congruence. Basis vectors were obtained from logarithmically amplitude-coded normalized band energies.

scaling on the analysis. Normalization, if used, was accomplished by scaling frames of spectral data so that the sum of the amplitude-coded band energies would be constant in each frame. Presumably, better results would be obtained using normalization if the perceptual process tends to amplitude normalize prior to extracting other information features from speech.

In order to satisfy the assumption that statistical variance is a good measure of "information," each data component should have perceptual importance proportional to variance. Therefore all results reported in this study were derived from high-frequency preemphasized speech (6 dB/octave up to 3000 Hz), since this preemphasis approximates the equal subjective intensity contour of hearing (Stevens, 1972). In all cases, the silent portions of the speech material were excluded from the statistical analysis by using a threshold to exclude all frames having less than about −40 dB of total frame energy relative to the loudest speech sections. A total of 20 band energies were uniformly spaced on the perceptual frequency scale of mels since this spacing implies that each band energy will make an approximately equal contribution to the articulation index, a measure of perceptual importance (French and Steinberg, 1947). The empirical relationship

$$m = 2595 \log_{10}(1+f/700)\,,$$

was used to relate the frequency in mels, m, to the frequency in Hz, f, (Makhoul and Cosell, 1976). Center frequencies for the 20 band energies used in this study ranged from 400 mels (298 Hz) to 2300 mels (4454 Hz), with each filter having a 100-mel bandwidth, approximately the width of one critical band (Zwicker, 1961).

B. Speech analysis and synthesis

The speech samples analyzed in this study were recorded in a low noise environment after the high-frequency preemphasis below 3000 Hz and low-pass filtering above 5000 Hz at 36 dB/octave (6 pole Butterworth). Recordings were made of five adult male speakers and five adult female speakers, each reading the phonetically balanced "Rainbow Passage" (Snidecor and Isshiki, 1965). Each reading lasted about 50 to 60 s at an average reading rate. Both Li *et al.* (1969) and our own experiments showed that the values of the covariance matrix of speech spectral band energies, and thus the principal-components basis vectors, stabilize after about 30 s of speech.

All further analysis was performed digitally on a PDP-15/20 18 bit minicomputer. Transfer of one-quarter speed analog signals to the computer was accomplished with a 9 bit A-to-D converter at a sampling rate of 2.5 kHz (10 kHz real time). Digital data was stored on computer DEC tape (about 30 s of speech per DEC tape) for later processing. Calculations were done with floating point arithmetic with no attempt at real-time processing. Digitized samples of synthetic speech were transferred to a tape recorder at a one-quarter real-time rate.

Figure 4 is a block diagram of the overall speech analysis-synthesis system implemented on the computer. All analysis was performed on 20.0-ms overlapping Hamming-weighted data sequences spaced 12.8 ms apart. Band energies were computed from LP smoothed spectra for some of the pilot experiments, but computed directly from FFT obtained spectra for the data reported in this paper. Prior to computing band energies from the 256 point FFT's, each power spectral point was averaged over five FFT values (about 156 Hz) which, in addition to the smoothing caused by the 20.0-ms time window, caused the skirts of each simulated band-pass filter to overlap adjacent filters by about 88 Hz. For each of the amplitude codings mentioned above and each speaker, the covariance matrix and its eigenvectors were computed. For the two speaker groups (males and females), group-averaged basis vector sets were calculated for each amplitude coding.

Speech synthesis was performed using a combination spectral principal-components LP vocoder, so that information retained by various numbers of principal components for various methods could be tested. For each principal-components vocoder, principal components were calculated from amplitude-coded band en-

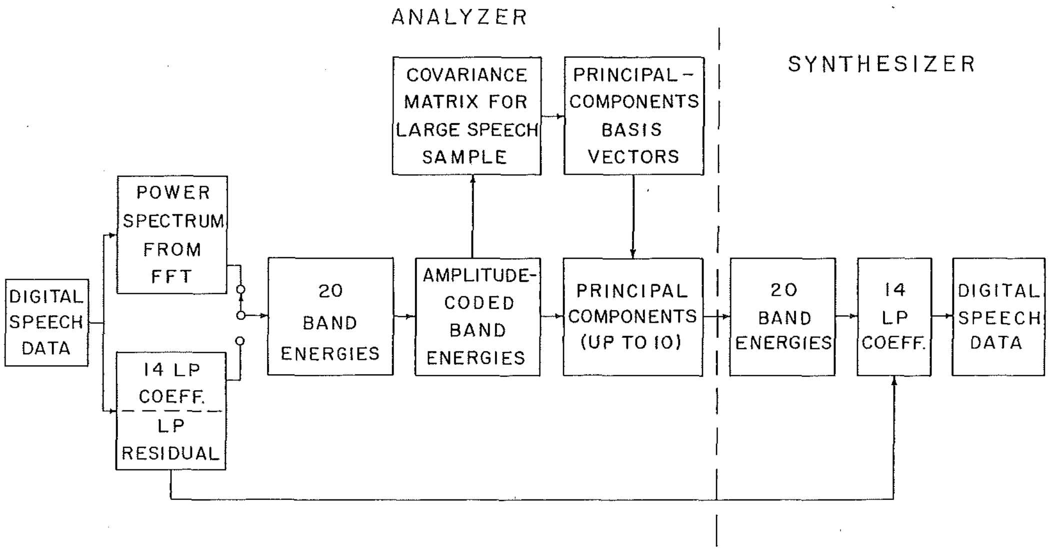

FIG. 4. Speech analysis-synthesis system.

ergies and the appropriate group-averaged basis vectors. The 20 band energies were reestimated for synthesis purposes, using the procedure indicated in Fig. 1. Fourteen LP coefficients were computed from the reestimated band energies for use by the LP synthesizer. The 14th order LP model was used to insure that spectral degradation was due almost entirely to the principal-components data reduction, and not to the LP spectral modeling of principal-components derived spectra.

The LP coefficients were computed from the band energies by first calculating autocorrelation coefficients using the formula:

$$R(i)=\sum_{n=0}^{20} P(f_n)\Delta f_n \overline{\cos(i2\pi f_n)}, \quad \text{for } i=0,1,\ldots p,$$

where

$R(i)=i$th autocorrelation coefficient, $P(f_n)=$ band level of the nth band, $\Delta f_n=$ bandwidth of the nth band, $\overline{\cos(i2\pi f_n)}=$ average of $\cos(i2\pi f_n)$ over the nth band, and $p=$ number of LP coefficients to be calculated.

This formula accounts for both the nonuniform bandwidths of the various filters, and the nonuniform spacing (on a linear scale) of the filter center frequencies. $P(f_0)$, representing the filter from DC to 215 Hz, was experimentally found to approximately equal $0.2P(f_1)$, and thus $0.2P(f_1)$ was used for $P(f_0)$ in calculating the autocorrelation coefficients. The filter from DC to 225 Hz was not used in the statistical analysis because of its very low contribution to the articulation index. Durbin's recursive procedure (Makhoul, 1975) was used to determine the LP coefficients from the autocorrelation coefficients.

In order to avoid problems of pitch detection and estimation that frequently plague vocoders (for example, McGonegal *et al.*, 1977), the LP residual was calculated in the analysis stage of the vocoder and used for vocoder excitation in the synthesizer. A 14th order LP inverse filter, rather than the more customary 10 or 12 coefficient filter, was used to minimize the information content of the residual signal. Both listening tests and examinations of spectrograms indicated that the LP residual contained only minimal spectral information. Attempts to further reduce remanent spectral information in the LP residual by center clipping produced little noticeable difference in the auditory quality of the residual signal, but did introduce an annoying harshness in resultant synthetic speech.

Prior to incorporation of the principal components into the vocoder, the basic LP vocoder was tested and optimized. Initial testing indicated that a 20.0-ms analysis frame resulted in somewhat higher quality synthetic speech than either a 17.0- or 25.6-ms analysis frame; therefore the 20.0-ms analysis frame was used for both the vocoder and the FFT analysis frame length. Speech synthesized from the basic LP vocoder (14 LP coefficients, autocorrelation method, 20.0-ms analysis frame, 12.8-ms frame update rate, residual excited) was found to be almost indistinguishable from the original speech. Speech synthesized using 14 LP coefficients derived from 20 band energies (that is, no principal-components data reduction) was found to be extremely high in quality, but somewhat inferior to the basic 14 pole LP vocoder, as described above.

Two types of control vocoders were used for comparison with principal-components vocoders. One type of control vocoder was an LP vocoder with the same number of parameters as the corresponding test principal-components vocoder. The other control vocoder used was a linear predictive spectrally warped (LPCW) vocoder, which is an LP vocoder which matches the LP model spectrum to the speech spectrum more closely at low frequencies than at high frequencies (Makhoul and Cosell, 1976). Except for the spectral warping property of the LPCW vocoder, the LP and LPCW vocoders were the same. For the control vocoders, energy was counted as one parameter since the signal energy is included in the principal components. Thus, for example, a four parameter LP vocoder has three LP coefficients plus signal energy as the fourth parameter.

All vocoders, both principal-components and control, were identical except for the method used to encode the spectral information. For the principal-components vocoders, spectral information was encoded by the principal components. For the LP and LPCW control vocoders, spectral information was encoded in terms of the LP coefficients. The analysis frame time, frame update rate, and excitation signal were the same for all vocoders. Except for one experiment in which white noise was used as the excitation, the LP residual signal (obtained from a 14 pole LP inverse filter in all cases) was used as the excitation signal.

C. Intelligibility testing procedures

The ultimate criterion for evaluating principal-components techniques for use in encoding speech spectra is the amount of speech information which is retained by a given number of components and/or the data rate of those components. We felt that the most practical method for measuring this information was to measure the intelligibility and quality of speech synthesized from the principal components. In the present study, we have characterized our data-reduction methods in terms of the number of "slowing varying" parameters rather than the data rate of those parameters, although, presumably, data rate is closely related to the number of parameters. Moreover, for speech preprocessing in certain fields (such as sensory substitution for the deaf), characterization of speech compression systems in terms of the number of parameters may be more useful than characterization in terms of data rate.

A form of the Diagnostic Rhyme Test (DRT) developed by Voiers *et al.* (1973) was used for evaluating speech intelligibility. The task of the listener in the DRT is to distinguish between minimally contrasting rhyming words of a word pair. In our test, we used a subset of 30 word pairs from the DRT (five word pairs from each of the six feature categories included in the DRT), plus ten additional word pairs. The added word pairs, all CVC words, contrast vowels closely spaced in the $F1$–$F2$ vowel plane rather than initial consonants, as do all the word pairs in the standard DRT. The decision to use strictly consonant pairs in the DRT is based on the observation that the bulk of the information in English is carried by the consonants. However, we added the contrasting vowel pairs to the test, since an intelligible and natural sounding speech system should faithfully transmit the vowel sounds.

The actual word pairs used for testing are given in Table I. From the word pairs, randomized word lists were made with the first or second word of a pair randomly selected. These randomized word lists were processed by each vocoder. A panel of listeners evaluated the vocoder output speech, hearing blocks of 20 words from a particular vocoder. The twenty-word blocks were randomized among the various vocoders and listeners to minimize effects due to training, boredom, or fatigue.

Since the acceptability of the output of a voice communication system can be influenced by factors other than intelligibility (Voiers, 1977), we also used an A/B paired sentence preference test for evaluating the quality of our principal-components vocoders. For this sentence test, listeners were instructed to select the sentence of a pair (identical sentences except for processing method) which they believed to be more "natural sounding," without particular regard to intelligibility. The following seven sentences were chosen for use in the sentence preference experiment because they are representative of a large variety of speech events, are fairly short, and also have been used in previous similar experiments (McGonegal *et al.*, 1977; Huggins *et al.*, 1977):

(1) We were away a year ago.

(2) I know when my lawyer is due.

(3) Every salt breeze comes from the sea.

(4) I was stunned by the beauty of the view.

(5) His vicious father has seizures.

(6) The little blankets lay around on the floor.

(7) The trouble with swimming is that you can drown.

TABLE I. Word pairs for modified DRT intelligibility test.

1.	Veal–Feel	21.	Weed–Reed
2.	Bee–Vee	22.	Tea–Key
3.	Meat–Beat	23.	Heed–Hid
4.	Keep–Cheep	24.	Cut–Cot
5.	Dune–Tune	25.	Pool–Tool
6.	Choose–Shoes	26.	Rue–You
7.	News–Dues	27.	Hid–Head
8.	Goose–Juice	28.	Caught–Cot
9.	Zed–Said	29.	Met–Net
10.	Den–Then	30.	Wren–Yen
11.	Mend–Bend	31.	Head–Had
12.	Care–Chair	32.	Bud–Bird
13.	Daunt–Taunt	33.	Bong–Dong
14.	Chaw–Shaw	34.	Taught–Caught
15.	Gnaw–Daw	35.	Pad–Pod
16.	Gauze–Jaws	36.	Hood–Heard
17.	Bond–Pond	37.	Wad–Rod
18.	Bon–Von	38.	Dot–Got
19.	Mom–Bomb	39.	Caught–Cut
20.	Cop–Chop	40.	Should–Shoed

III. SPEECH SPECTRAL DATA

Figure 5 shows the spectral mean values and variances for the logarithmically amplitude-coded non-normalized spectral band energies. The data is normalized and displayed as percent of the total. The spectral means tend to be relatively constant versus frequency for both speaker groups except for rather broad peaks around 500 and 2200 Hz. The spectral variances depend on frequency to a larger degree than do the spectral mean values, but are almost always between 2% and 8%. Thus if any band energies were deleted from a channel vocoder speech synthesizer, between 2% and 8% of the total variance would be deleted. The between-speaker differences for the variances are the largest in regions corresponding roughly to the formant

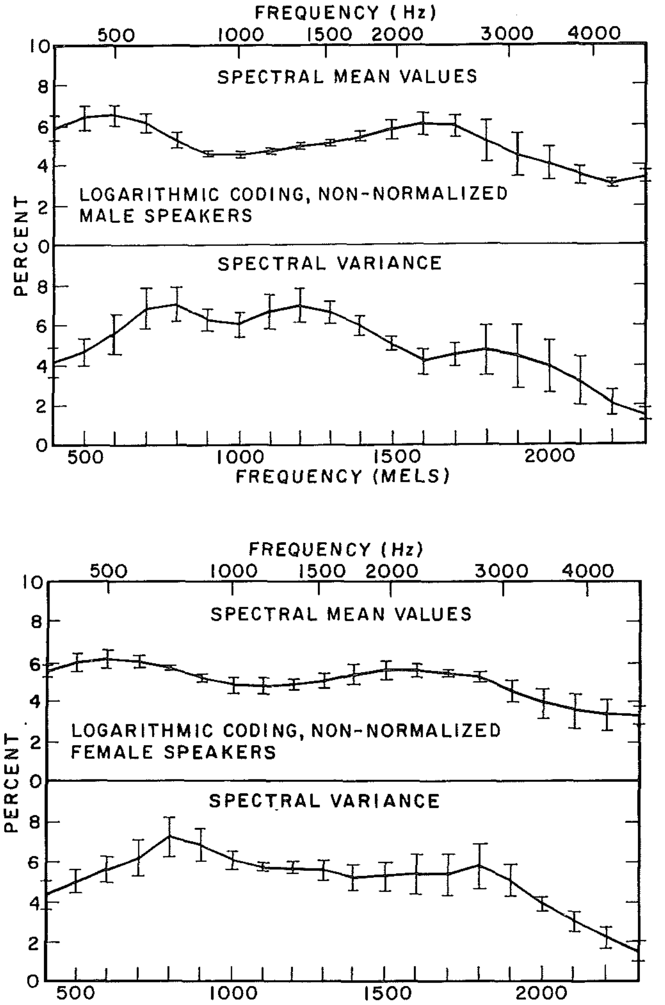

FIG. 5. Mean values and variance of logarithmically amplitude-coded non-normalized spectral band energies. Data are the average from five speakers for each curve. Vertical bars represent two standard deviations.

locations near 600, 1500, and 3000 Hz for the male speakers, and near 400–800 Hz and near 1800–2500 Hz for the female speakers. The spectral mean-value and variance plots obtained from the other amplitude codings used in this study are similar to the ones shown in Fig. 5.

Plots of group-average cumulative variance as a function of the number of principal-components basis vectors are shown in Fig. 6. These data show that the speech spectral band energies are highly correlated, since a very high percentage of total spectral variance can be accounted for with a small number of dimensions. For example, the first five principal components contain about 90% of the total variance. Plots were also made of cumulative variance versus number of principal-components dimensions for the other nonlinear amplitude codings investigated in this study in order to obtain preliminary comparisons of the performance of the various amplitude scales. If these various amplitude codings were to be ranked in terms of most cumulative variance for a given number of dimensions, the results for both the female and male groups would be normalized logarithmic (coding B), non-normalized logarithmic (coding A), non-normalized $\frac{1}{3}$ power function (coding D), normalized $\frac{1}{3}$ power function (coding C). For example, using five dimensions, coding B accounts for about 92% of the variance and coding C about 87% of the variance. These rankings are based strictly on maximum variance for a given number of dimensions and are, in fact, different from those obtained from the speech synthesis experiments.

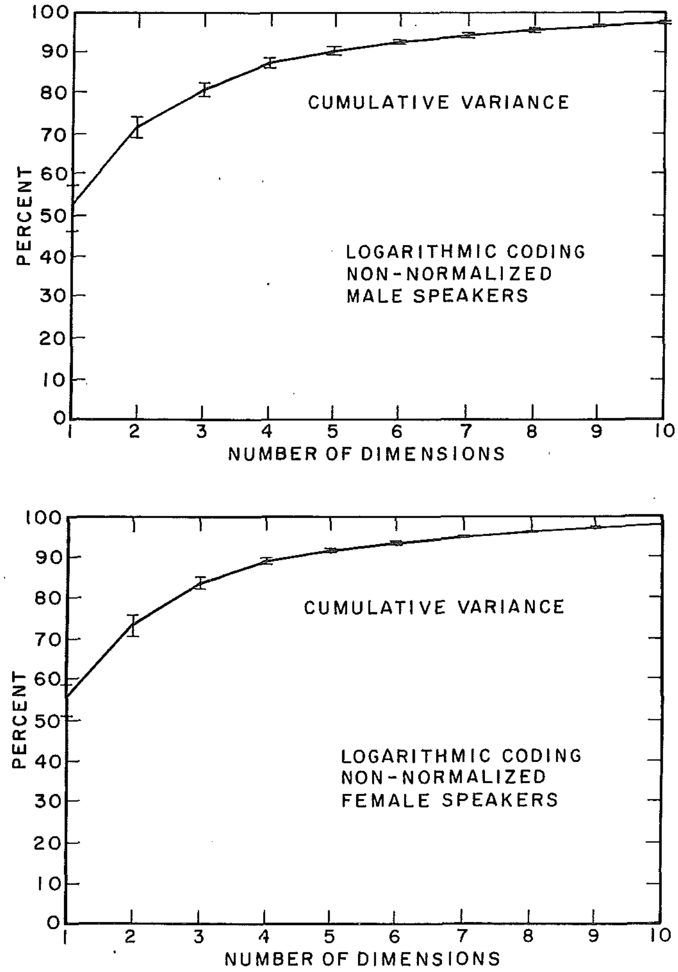

FIG. 6. Cumulative variance as a function of number of dimensions (principal components) for logarithmically amplitude-coded non-normalized spectral band energies. Data are the average from five speakers for each curve. Vertical bars represent two standard deviations.

The eigenvectors for the various speakers were obtained individually and orthogonally rotated to congruence before averages and between-speaker differences were computed. However, the eigenvectors obtained for the various individual speakers tended to be fairly similar even prior to rotation to congruence. In particular, the first four or five basis vectors were very similar. A practical problem associated with the orthogonal rotation to congruence procedure is that the rotation merely rotates the basis vectors of one set toward those of another set and is not a general technique for optimally rotating several basis vector sets toward common congruence. Therefore we adopted the somewhat *ad hoc* procedure of choosing a typical speaker for each group whose eigenvectors appeared to be most representative of the speakers within that group and rotated the eigenvectors of all the other speakers within the group toward those of the typical speaker. The first four group-averaged basis vectors are shown in Fig. 7 for the male speakers for the non-normalized log-coded speech spectral data. The corresponding basis vectors for the female speakers are shown in Fig. 8.

Very roughly speaking, these basis vector sets are similar to a Fourier series basis vector set. This type of basis vector set was theoretically predicted by Yilmaz (1967) for speech spectra encoded with perceptual amplitude and frequency scales and also is similar to the eigenvector set obtained in the experimental study by Li *et al.* (1969). The first basis vector is roughly constant as a function of frequency and thus the first principal component will be a measure of energy. The second basis vector, as a function of frequency, is similar to one negative cycle of a sinusoid from about 300 to about 4500 Hz with a crossover at about 1500 Hz. The second principal component, together with the first principal component, will be a measure of the spectral mean. Therefore the second principal component will be an indication of whether the spectrum is more heavily weighted below 1500 Hz (for example, most vowels) or more heavily weighted above 1500 Hz (for example, most fricatives and consonants), and will help separate those vowels having a high $F2$ from those having a low $F2$. The third and fourth basis vectors give information about increasingly specific parts of the spectrum. For example, eigenvector three is most heavily peaked between 300 and 500 Hz (slightly lower than the most common first formant frequencies), whereas eigenvector four tends to be most heavily peaked in the second formant range of 1500 to 2000 Hz.

Basis vectors obtained from normalized spectra tend to be similar to the basis vectors for the non-normalized spectra, described above, except displaced by one

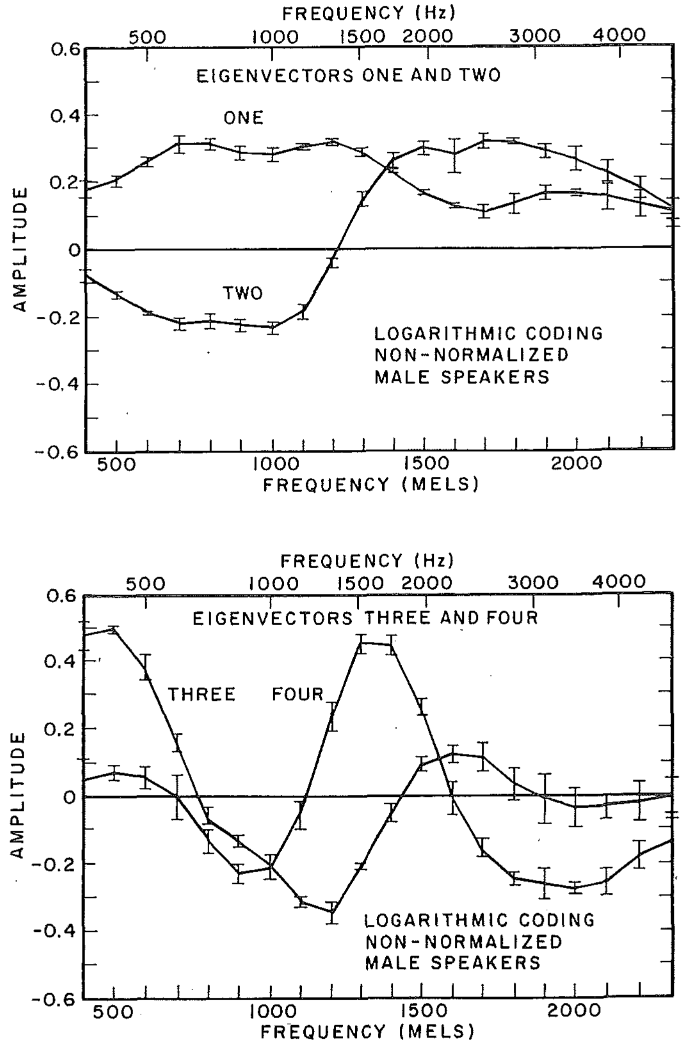

FIG. 7. First four principal-components basis vectors obtained from analyzing logarithmically amplitude-coded non-normalized spectral band energies. Data are the average from five male speakers. Vertical bars represent two standard deviations.

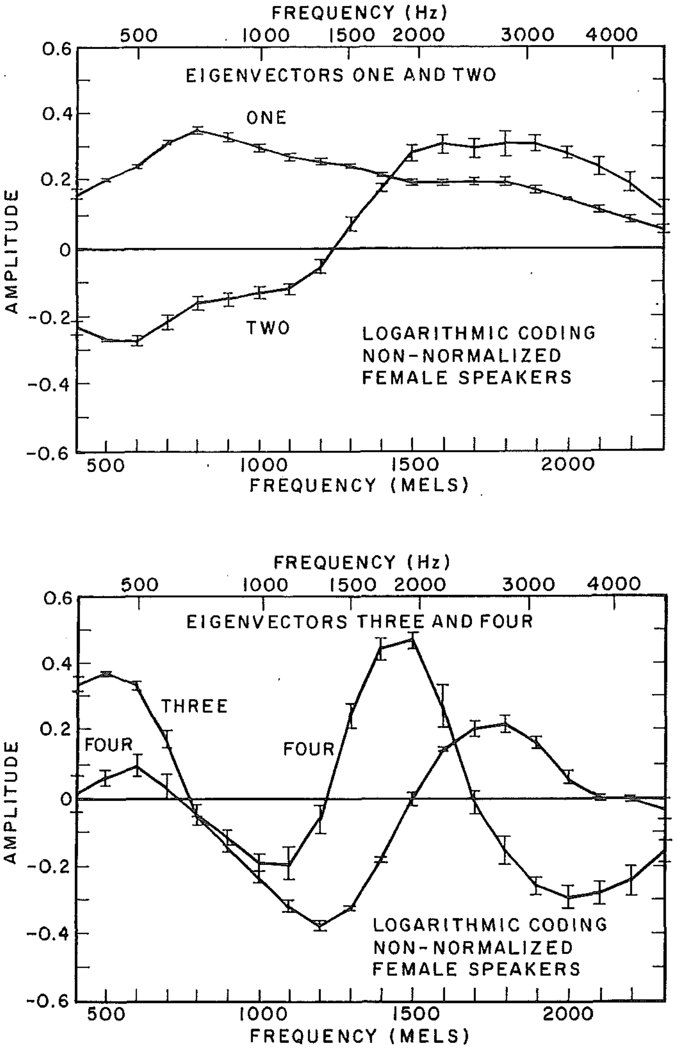

FIG. 8. First four principal-components basis vectors obtained from analyzing logarithmically amplitude-coded non-normalized spectral band energies. Data are the average from five female speakers. Vertical bars represent two standard deviations.

number. That is, for the normalized spectra, there is no constant vector, basis vector one is similar to the second basis vector for the non-normalized spectra, and so on. For the higher numbered vectors, the similarities between corresponding (displaced by one number) basis vectors of normalized and non-normalized spectra are less than for the smaller numbered vectors.

Basis vectors obtained from $\frac{1}{3}$-power-function-coded spectra are similar to the basis vectors shown in Fig. 7 and Fig. 8. However, the basis vectors obtained from power-function-coded spectra are generally less smooth than those from the log-coded spectra. Also, the between-speaker differences are usually larger for the power-function coding than the log coding.

Comparison of Fig. 7 and Fig. 8 shows that the basis vectors for males and females are quite similar for the first two basis vectors. For basis vector two, even the crossover frequencies are very close. However, the higher-ordered basis vectors for the female speakers are shifted noticeably higher in frequency than for the corresponding male basis vectors. For example, the dominant peak in basis vector four is about 100 mels higher (1900 versus 1600 Hz) in frequency for the female group than for the male group. This frequency shift corresponds approximately to the average difference between the second formant frequencies of male and female speech. For our limited group, it is also seen that the within group variability is less for the females than the males.

IV. SPEECH INTELLIGIBILITY AND SPEECH QUALITY EXPERIMENTS

In this section, we present the results of speech synthesis experiments conducted to evaluate the intelligibility and quality of speech synthesized from spectral principal components. For all the speech synthesis experiments, speech systems were characterized in terms of the number of spectral parameters and not by the bandwidth required to transmit those parameters. Therefore all principal components (for the experimental principal-components vocoders) and LP coefficients (for the control vocoders) were updated at the same rate as the vocoder analysis frame update rate (once every 12.8 ms) with full floating point precision, and no attempt was made to determine the effect of parameter bandwidth on intelligibility.

Based on the results of informal listening experiments, we concluded four parameter systems were the most useful for evaluating differences among the principal-components variables of this study (spectral band-energy amplitude scales). Some reasons are (1) speech synthesized from any of the three parameter systems seemed to be marginal in both quality and intelligibility. (2) There seemed to be substantial overall improvement between three parameter and four parameter systems. (3) Within the four parameter systems, there appeared to be fairly large differences (at least in terms of speech quality) among the different systems. (4) The overall improvement between four and five parameter systems was much less than the improvement achieved by changing to four parameters from three parameters. (5) The informal tests also indicated that the differences in intelligibility scores for speech synthesized from five or more principal components would be relatively small. Since the four parameter systems were tested more extensively than the other systems, confidence intervals, indicating plus or minus one standard deviation, are indicated only for the test results of the four parameter systems.

Informal listening tests also indicated that speech synthesized from principal components based on linear spectral amplitude coding was substantially poorer in quality and intelligibility than the speech synthesized from principal-components systems based on the nonlinear scales investigated in this study. Therefore the linear amplitude scaling was not included in the more complete intelligibility tests reported in this section.

In total, five speakers (three males and two females) were used for the synthesis experiments reported in this paper. None of the speakers used in the intelligibility experiments had been used in the earlier statistical analysis experiments; however, two of the speakers used for the sentence preference experiment had been used as subjects for the statistical analysis experiments. All speech synthesis was performed using group-averaged principal-components basis vectors. The only speaker-dependent terms used in speech synthesis were the average value terms, i.e., the M_j from Fig. 1. The M_j terms represent a rather complex but nontime-varying filter. This filter function was calculated separately for each speaker, using about 10 s of speech, since informal listening tests indicated slight losses in quality when only one filter was used for all speakers. However, because this filter is nontime-varying for each speaker, the information rate to specify the parameters of the filter is negligible.

The listening crews for each of these experiments consisted of eight to ten young adults—about half males and half females. The listeners were not given any training for the experiments, and the scores reported were obtained from listening to all the test materials once. About half the listeners participated in either two or three of the experiments and therefore were somewhat "experienced" in the later experiments. For all listening experiments, volume levels were adjusted to a conversationally comfortable level. The test materials were presented binaurally over headphones in a room with relatively little background noise.

A. Intelligibility experiment 1

This test was performed for one male speaker with a panel of eight listeners using the modified DRT discussed above with the 80 words from Table I. The intelligibility scores for principal-components and control vocoders of three, four, and five channels are depicted in Fig. 9. All intelligibility scores are given in terms of percent correct, after adjustment for the effects of chance using

$$P_c=(R-W)/T\,,$$

where

P_c= percent correct, after adjustment for chance, R= number of correct responses, W= number of incorrect responses, T= total number of responses.

From Fig. 9 we see that the five parameter principal-components systems are about 80% intelligible, the four parameter systems about 75% intelligible, and the three parameter system about 62% intelligible. The only principal-components system which appears to be significantly poorer than the others is system D (non-normalized $\frac{1}{3}$ power-function coding). For the four parameter systems, the hypothesis that system D is the poorest principal-components vocoder and that the LP vocoder is worse than the LPCW vocoder, can be accepted at the 95% confidence level. The intelligi-

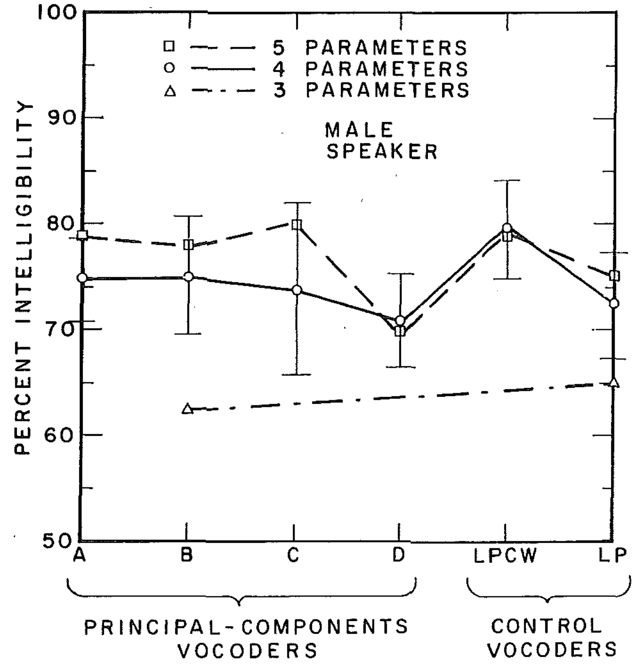

FIG. 9. Modified DRT speech intelligibility scores for various principal-components and control vocoders. Vocoder excitation was the LP residual in all cases. Vertical bars represent two standard deviations for the four parameter systems.

bility scores for the LPCW control vocoder were somewhat higher than those for the LP control vocoder for both four and five parameters. Other results obtained from intelligibility experiment 1 are

(1) Contrasting vowel pairs were typically easier to differentiate than contrasting consonant pairs (vowel scores typically about 10% higher than consonant scores).

(2) The intelligibility scores for the LP residual averaged about 14%.

(3) The intelligibility scores for both the original speech and vocoded speech obtained with a 20 channel vocoder averaged about 90%. This result indicates that many of the word distinctions, as pronounced by this speaker, were apparently less clear than optimum and therefore this speaker was not used in other experiments.

B. Sentence preference experiment

Altogether, four speakers and seven sentences were used in the A/B paired sentence preference test, described in Sec. IIC to test the speech quality of various principal-components vocoders. The ten subjects participating in the experiment were asked to select that sentence of a pair which "sounded better," using their personal evaluation criteria. In the experiment, a total of 160 preference judgements were made by each subject. All sentence pairs were used twice with the order of repetition interchanged to eliminate possible subject biases toward selecting the first or second sentence of a pair. About 2 s were allowed between sentences of a pair and about 4 s between sentence pairs.

The largest percentage of the comparisons were between those three, four, and five channel principal-components vocoders which appeared to be the best (A, B, and C). Type D principal-components vocoder (non-normalized $\frac{1}{3}$ power-function coding) was not tested as extensively, since both the first intelligibiltiy experiment and informal listening experiments rated this vocoder to be uniformly poorest in performance. Some of the sentence pairs were also used to make comparisons between principal-components and control vocoders.

The test sentences for the preference experiment were also used to obtain an estimate of word intelligibility within a sentence context. Prior to making the sentence comparisons, the subjects listened to one repetition of each sentence, synthesized from four principal components corresponding to method B, and were asked to record each sentence. On the average, the subjects correctly identified about 95% of the 51 words in the seven test sentences.

The primary results of the sentence preference experiments are depicted in Fig. 10. Results are given in terms of mean percent preference, with averages taken as follows. Principal-components vocoders A, B, and C were compared against each other and the mean preferences are based on these comparisons. Principal-components vocoder D was compared with vocoders B and C, and the score indicated is the average of these two scores. The LPCW control vocoder score is strictly based on a comparison with principal-components vocoder A. For the results shown in Fig. 10, comparisions were made only among vocoders with the same number of parameters. Not all systems were compared with all other systems because of the large amount of testing that would have been required.

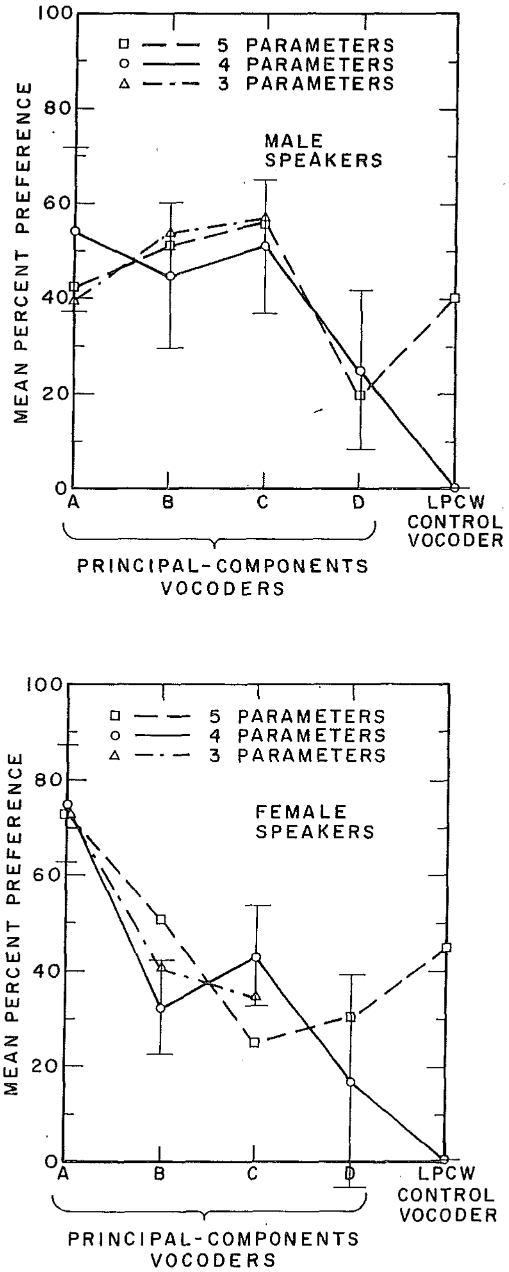

FIG. 10. Sentence preference ratings for various principal-components vocoders and the LPCW control vocoder. All data shown are for comparisons of sentences synthesized from the same number of parameters. Vertical bars represent two standard deviations for the four parameter systems.

The data for the female speakers in Fig. 10 show that among the principal-components systems with highest intelligibility (A, B, and C), system A (non-normalized log coding) is a strong favorite. Next in order of preference are systems B and C with fairly similar ratings. System D, with the worst intelligibility rating, also had a preference rating substantially worse than for the other systems. Note that system D averaged only about 25% preference even though it was only compared with the third and fourth place systems B and C. The LPCW control vocoder has a substantially lower rating than the best principal-components vocoder, especially for four parameters, for which the control vocoder was never preferred by any of the subjects.

The data from the male speakers in Fig. 10 show that systems A, B, and C all receive about the same ratings, and only system D appears to be clearly less preferred. If, however, we continue to argue that the four parameter data is the most significant, principal-components vocoder A is a slight favorite. The LPCW control vocoder again has a lower rating than principal-components vocoder A, especially for four parameters. For both the male and female speakers, the majority of the data obtained from the various listeners was within about ±15% of the mean preference for all listeners, as indicated by the confidence intervals for the four parameter systems.

C. Intelligibility experiment 2

An additional intelligibility experiment was performed using one male and one female speaker with the 80 words from Table I. This experiment tested only three and four parameter versions of principal-components systems A, B, and C and the LPCW vocoder. This intelligibility experiment differed from all the other speech synthesis experiments in that the vocoder sound source in this experiment was Gaussian band-limited (0 to 5 kHz) noise. This sound source was chosen so that the synthetic speech information would be entirely derived from the spectral parameters, and so that the synthetic speech intelligibility scores would (hopefully) be somewhat lower than corresponding scores obtained from an LP residual-excited vocoder and more sensitive to the information contained in the spectral parameters. This particular artificial sound source was chosen because it is similar to the human sound source for whispered speech; thus the synthetic speech in this experiment sounded somewhat like a hoarse whisper. The original speech material for this test was found to be about 99% intelligible in another independent experiment.

Results of the experiment, in terms of percent correct after adjustment for the effects of chance, are shown in Fig. 11 for one male speaker and one female speaker. The scores are about 80% for the four parameter systems and about 70% for the three parameter systems. In this experiment, there seems to be no universal clearcut preference among the various principal-components systems. For the case of four parameters, however, the hypothesis that all the principal-components vocoders are better than the LPCW vocoder can be accepted at the 95% confidence level. The scores for the female speaker are also universally higher than the corresponding scores for the male speaker. In spite of the absence of a "natural" sound source for the vocoders used in this experiment and the use of untrained listeners as subjects, the average intelligibility scores shown in Fig. 11 are somewhat higher than those shown in Fig. 9, presumably due to the much greater clarity of speech for the speakers in this experiment, compared to the speaker used for Fig. 9.

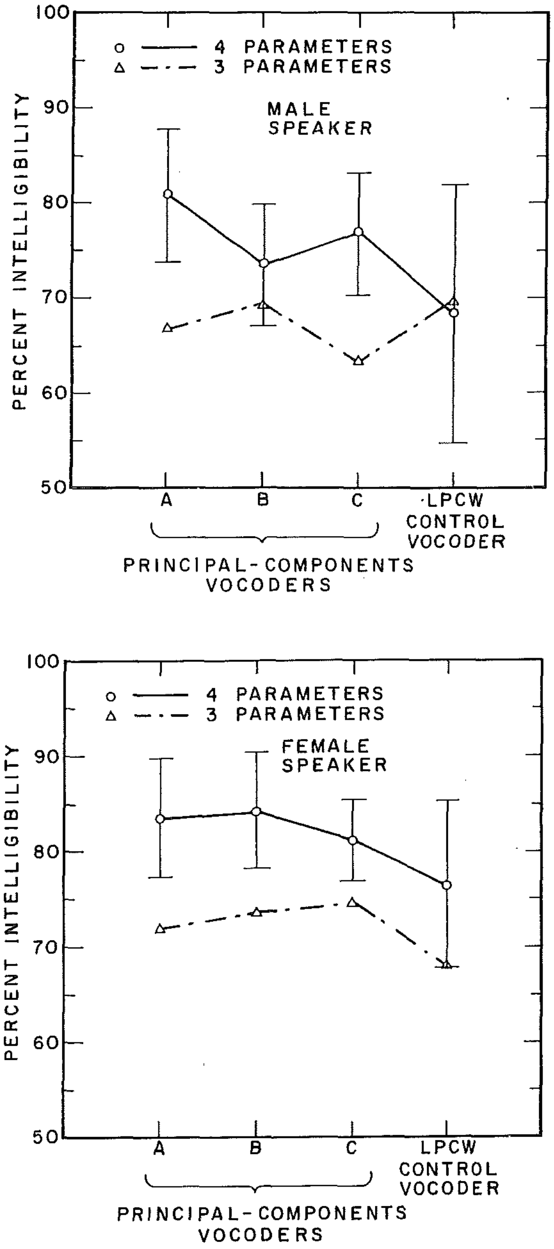

FIG. 11. Modified DRT speech intelligibility scores for various principal-components vocoders and the LPCW control vocoder. Band-limited Gaussian noise was used for the vocoder excitation in all cases. Vertical bars represent two standard deviations for the four parameter systems.

V. CONCLUSIONS

The results of this study indicate that it is possible to encode a very high percentage of speech information with as few as three to five spectral principal components. The intelligibility of speech synthesized from principal components, based on modified DRT intelligibility scores, can be summarized as follows: Speech synthesized from three principal-components spectral parameters is about 70% intelligible, speech synthesized from four parameters is about 80% intelligible, and speech obtained from five parameters is about 85% intelligible. Speech synthesized from the best principal-components vocoders is at least as intelligible, or perhaps slightly more intelligible than speech synthesized from LP and LPCW vocoders with the same number of parameters. The exact interpretation of these intelligibility scores obtained within the restricted framework of the DRT is unknown. However, we did find that about 95% of the words in unfamiliar short sentences were correctly identified for the four parameter systems (80% intelligibility score with the DRT).

Principal-components basis vectors obtained from speakers of the same sex are very similar. Basis vectors derived from female speech are fairly similar to those derived from male speech except that some of the female basis vectors are shifted toward higher frequencies relative to the basis vectors of male speakers. In general, the principal-components basis vectors corresponding to the largest covariance matrix eigenvalues are largely speaker independent, whereas the basis vectors corresponding to the smallest covariance matrix eigenvalues vary most from speaker to speaker.

The particular spectral amplitude measurement scale used with a principal-components analysis is not critically important provided at least some nonlinear scale compression is used. However, of the scales tested in this study, based on the combined results of intelligibility tests, a sentence preference test, and informal listening tests, the order of preference for speech spectral amplitude scales is (1) non-normalized logarithmic (coding A), (2) normalized logarithmic (coding B), (3) normalized $\frac{1}{3}$ power function (coding C), (4) non-normalized $\frac{1}{3}$ power function (coding D), and (5) non-normalized linear (coding E). The differences between codings A, B, and C are fairly small, whereas the difference between coding C and D is larger, and the difference between D and E is quite large.

In summary, our study indicates that the linear scale is clearly the poorest as a measurement scale for speech spectra if mean-square error is used to measure differences between spectra, and that the dB scale is slightly preferred over a sone scale. The preference for the dB scale over the sone scale may seem surprising in that the sone scale is the psychophysical ratio scale for loudness. However, we believe the preference for the dB scale over the sone scale arises from the fact that jnd's increase in magnitude on the sone scale for larger sone values whereas jnd's are roughly a constant number of dB for various speech levels. Thus if we make the assumption that the perceptual distance between two fairly similar speech sounds is more related to the total number of jnd's separating the spectra than their absolute difference on a perceptual ratio scale, then the dB scale is to be preferred over the sone scale. To compare with a simple example, the scale of centimeters is approximately a perceptual ratio scale of distance. However, in describing the differences between the lengths of two lines, the percentage differences are probably more important perceptually than the absolute differences in centimeters.

ACKNOWLEDGMENTS

This work was supported in part by the Rome Air Development Center under the Air Force Post-Doctoral Program, Contract No. F30602-75-C-0121, and in part from research Grant ENG 77-25003 from the National Science Foundation.

Atal, B. S., and Hanauer, S. L. (1971). "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," J. Acoust. Soc. Am. 50, 637–655.

Boehm, J. F., and Wright, R. D. (1968). "Dimensional Analysis and Display of Speech Spectra," J. Acoust. Soc. Am. 44, 386.

Cliff, N. (1966). "Orthogonal Rotation to Congruence," Psychometrika 31, 33–42.

Crowther, W. R., and Rader, C. M. (1966). "Efficient Coding of Vocoder Channel Signals Using Linear Transformations," Proc. IEEE 54, 1594–1595.

Fechner, G. T. (1860). *Elemente der Psychophysik* (Breitkopf & Hartel, Leipzig).

French, N. R., and Steinberg, J. C. (1947). "Factors Governing the Intelligibility of Speech Sounds," J. Acoust. Soc. Am. 19, 90–119.

Harman, H. H. (1976). *Modern Factor Analysis* (Chicago U. P., Chicago, London), pp. 133–164.

Huggins, A. W. F., Viswanathan, R., and Makhoul, J. (1977). "Speech-Quality Testing of Some Variable-Frame-Rate (VFR) Linear-Predictive (LPC) Vocoders," J. Acoust. Soc. Am. 62, 430–434.

Klein, W., Plomp, R., and Pols, L. C. W. (1970). "Vowel Spectra, Vowel Spaces, and Vowel Identification," J. Acoust. Soc. Am. 48, 999–1009.

Kramer, H. P., and Mathews, M. V. (1956). "A Linear Coding for Transmitting a Set of Correlated Signals," IRE Trans. Inform. Theory, IT-2, 41–46.

Kulya, V. I. (1964). "Experimental Investigation of the Correlation Relations in the Speech Spectrum of Some Variants of an Orthogonal Vocoder," Telecommun. Radio Eng. (USSR) 18, 39–50.

Li, K. P., Hughes, G. W., and House, A. S. (1969). "Correlation Characteristics and Dimensionality of Speech Spectra," J. Acoust. Soc. Am. 46, 1019–1025.

Makhoul, J. (1975). "Linear Prediction: A Tutorial Review," Proc. IEEE 63, 561–580.

Makhoul, J., and Cosell, L. (1976). "LPCW: An LPC Vocoder with Linear Predictive Spectral Warping," Conf. Rec. 1976 IEEE Int. Conf. Acoust. Speech, Signal Process., 12–14 April April, 1976, Philadelphia, PA (Canterbury, Rome, New York), pp. 466–469.

McGonegal, C. A., Rabiner, L. R., and Rosenberg, A. E. (1977). "A Subjective Evaluation of Pitch Prediction Methods Using LPC Synthesized Speech," IEEE Trans. Acoust., Speech, Signal Process., ASSP-25, 221–229.

Nierop, D. J. P. J. van, Pols, L. C. W., and Plomp, R. (1973). "Frequency Analysis of Dutch Vowels from 25 Female Speakers," Acustica 29, 110–118.

Plomp, R., Pols, L. C. W., and van de Geer, J. P. (1967). "Dimensional Analysis of Vowel Spectra," J. Acoust. Soc. Am. 41, 707–712.

Pols, L. C. W., Kamp, L. J. Th. van der, and Plomp, R. (1969). "Perceptual and Physical Space of Vowel Sounds,"

J. Acoust. Soc. Am. 46, 458–467.

Pols, L. C. W. (1971). "Real-Time Recognition of Spoken Words," IEEE Trans. Comput. C-20, 972–978.

Pols, L. C. W., Tromp, H. R. C., and Plomp, R. (1973). "Frequency Analysis of Dutch Vowels from 50 Male Speakers," J. Acoust. Soc. Am. 53, 1093–1101.

Pols, L. C. W. (1975). "Intelligibility of Speech Resynthesized by Using a Dimensional Spectral Representation, in *Speech Communication*, Vol. 1 (Almqvist and Wiksell, Stockholm), pp. 87–96.

Pols, L. C. W. (1977). *Spectral Analysis and Identification of Dutch Vowels in Monosyllabic Words.* (Academische Pers B.V., Amsterdam), Ph.D. dissertation.

Sambur, M. R. (1975). "An Efficient Linear Prediction Vocoder," Bell Syst. Tech. J. 54, 1693–1723.

Schonemann, P. H. (1966). "A Generalized Solution of the Orthogonal Procrustes Problem," Psychometrika 31, 1–10.

Snidecor, J. C., and Isshiki, N. (1965). "Air Volume and Air Flow Relationships of Six Male Esophageal Speakers," J. Speech Hear. Disord. 30, 205–216.

Stevens, S. S. (1936).. "A Scale for the Measurement of a Psychological Magnitude: Loudness," Psychol. Rev. 43, 405–416.

Stevens, S. S. (1957). "On the Psychophysical Law," Psychol. Rev. 64, 153–181.

Stevens, S. S. (1972). "Perceived Level of Noise by Mark VII and Decibels (E)," J. Acoust. Soc. Am. 51, 575–601.

Voiers, W. D., Sharpley, A. D., and Helmsoth, C. J. (1973). "Research on Diagnostic Evaluation of Speech Intelligibility," final report, contract No. AF19628-70-C-0182, AFSC.

Voiers, W. D. (1977). "Diagnostic Acceptability Measure for Speech Communication Systems," *Conf. Rec.* 1977 IEEE *Int. Conf. Acoust., Speech, Signal Process.*, 9–11 May, 1977, Hartford, CT (IEEE Service Center, Piscataway, NJ), catalog No. 77CH1197–3, 204–207.

Yilmaz, H. (1967). "A Theory of Speech Perception," Math. Biophys. 29, 793–825.

Zwicker, E. (1961). "Subdivision of the Audible Frequency Range Into Critical Bands (Frequenzgruppen)," J. Acoust. Soc. Am. 33, 248.