Title

Transform Representation of the Spectra of Acoustic Speech Segments with Applications, Part 2: Speech Analysis, Synthesis and Coding

Journal

IEEE Transaction on Speech and Audio Processing, 1

Authors

Algazi, Ralph
Cadwell, C.
Irvine, D.
et al.

Publication Date

1993

Peer reviewed

Transform Representation of the Spectra of Acoustic Speech Segments with Applications—II: Speech Analysis, Synthesis, and Coding

V. Ralph Algazi, Kathy L. Brown, Michael. J. Ready, David H. Irvine, Christie L. Cadwell, and Sang Chung

Abstract— **In Part I of this paper, we introduced a new approach to the representation of the speech spectral envelope which makes use of the Karhunen-Lòeve transformation of acoustic subword segments. This new signal-dependent representation captures, with a few KL vectors and transform coefficients, the perceptually and phonetically important structure of the spectral envelope. In this second part, we study the application of this new representation to the analysis, synthesis, and coding of speech. We propose simple quantization and coding strategies for the KL representation vectors as well as for the resulting transform coefficients. The resulting technique is a variable rate encoding scheme which achieves good speech quality at an average rate of 3.5 kilobits per second.**

INTRODUCTION

In Part I of this paper, we proposed a new approach to the representation of speech which captures efficiently the perceptually and phonetically important structure and constituents of the speech spectral envelope over acoustic segments.

For a multiframe acoustic speech segment, we have seen that by the Karhunen-Lòeve (KL) transformation of the critical band vectors, a few KL coefficients are adequate to represent the original spectral envelope.

Thus, the approach holds promise for efficient analysis and synthesis, and therefore coding of speech. It is the purpose of this paper to examine in some detail this specific application of the representation of the spectral envelope of acoustic subwords.

Use of this representation requires that both the KL coefficients and the KL transformation itself be available during synthesis. Therefore, strategies for quantization and coding of the spectral envelope information must accommodate the KL transformation vectors along with the the KL coefficients.

Manuscript received May 30, 1989; revised June 8, 1992. This research was supported in part by Pacific Bell, Apple Computer, Hewlett Packard, Intel, Signal Science Inc., and the University of California program MICRO. The associate editor coordinating the review of this paper and approving it for publication was Dr. Mark A. Clements.

V. R. Algazi, K. L. Brown, and D. Irvine are with the Speech Research Laboratory, Center for Image Processing and Integrated Computing (CIPIC), University of California, Davis, CA.

M. J. Ready was with CIPIC, University of California, Davis. He is now with Applied Signal Technology (AST), Sunnyvale, CA.

C. Caldwell was with CIPIC, University of California, Davis. She is now with C-Cube Microsystems, Milpitas, CA.

S. Chung was with CIPIC, University of California, Davis. He is now with Nokia Mobile Phone Inc., San Diego, CA.

IEEE Log Number 9208580.

We also need an efficient excitation source model to complement the spectral envelope. We make use of a simple mixed excitation source. The goal of this work is principally to study the KL transformation as an alternate representation strategy for the spectral envelope.

The overall analysis-synthesis scheme that we are considering has been implemented in a coder, whose overall block diagram is shown in Fig. 1. The approach is an adaptive hybrid of waveform subband coding, because of the direct representation of the spectrum in the frequency domain, and source modeling, by the use of pitch information and mixed excitation.

We shall emphasize the concepts and algorithms for analysis and synthesis, discuss briefly the encoding issues, and illustrate the results by examples. Some aspects of our approach have been use in older work on subband vocoders. More recent studies with some similarity to our work include the work of Flanagan *et al.* [1] where a bank of filters was postulated, the issue of phase representation studied, and coding performance determined. Other work by Atal considers transforms to segment and represent the LPC coefficients for very low bit rate coding [2]. Another recent paper by Griffin and Lim, which describes a high quality, medium rate vocoder based on the direct encoding of the speech spectrum, has some similarity to our work [3].

The data used in our work consist of approximately 100 sentences from the TIMIT database, bandlimited to 4 kHz using a 10th-order Chebyshev filter and resampled at 10 kHz [4], [5]. The processing of the speech data was performed for frames of 256 samples overlapped by 128 samples using a Hanning window.

II. ANALYSIS AND SYNTHESIS

We draw the distinction between analysis-synthesis and coding. In analysis-synthesis, we are concerned with the dimension and complexity of the representation vectors, and the resulting quality of the reconstructed speech. In coding, where all vectors are quantized and encoded, and bits are counted, an additional degradation due to quantization occurs. The overall quality of synthesized speech depends on a number of factors, among which are good representations of the spectral envelope, an accurate pitch determination, and a suitable excitation.

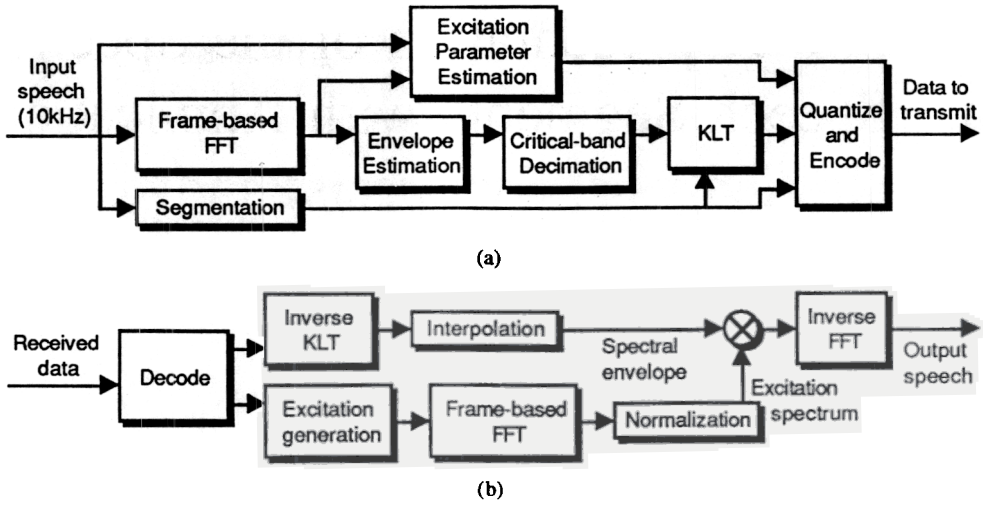


Fig. 1. Overall block diagram of the KL coder. (a) Analysis block diagram. (b) Synthesis block diagram.

Following our general speech representation approach, we make use of the critical-band representation of the spectral envelope. We adopted a pitch-synchronous mixed excitation as the excitation source in speech synthesis. The needed pitch information is extracted by a wavelet-based pitch detector. We begin this section with a brief treatment of the excitation model and pitch detection, followed by a detailed discussion of the spectral envelope representation.

A. Mixed Excitation

The excitation signal is well known to be critical to the quality of synthesized speech. The standard binary excitation model (quasiperiodic pulses for voiced speech, white noise for unvoiced speech) is incapable of producing natural-sounding speech: voiced speech synthesized with a binary model tends to sound buzzy and artificial. More elaborate excitation schemes where a "best" excitation signal is chosen from a large set of candidates (as in CELP, for example) are capable of much better-sounding speech but at far higher bit cost. We take an intermediate approach which uses a mixture of voiced and unvoiced energy during voiced speech with a single parameter specifying the mixture, and white noise for unvoiced speech.

Our excitation signal is based on the one described in [6]. In addition, we incorporate a flexible time-domain envelope for the unvoiced component during voiced speech. Without this envelope, the voiced and unvoiced components in synthesized voiced speech tend to remain perceptually distinct. However, with proper choice of the envelope shape, the voiced and unvoiced perceptual components merge into a single perceptual unit which is more natural-sounding. Estimation of the voicing mixture is performed by measuring deviation from periodic behavior in the spectrum of the original speech. Details of this excitation approach can be found in [7].

B. Pitch Extraction

To obtain pitch information to use in constructing the excitation signal, we use a pitch detector based on the one described in [8], [9], [10]. The speech signal is filtered by convolution with an appropriately scaled wavelet function ϕ', which is the derivative of a zero-phase low-pass smoothing function ϕ. The scale of the wavelet is chosen so that the energy in the wavelet-filtered speech is dominated by the fundamental frequency component. All peaks in the filtered signal are regarded as potential pitch events. (Peaks in the wavelet-filtered speech signal correspond to maxima in the derivative of the speech signal after it has been smoothed by convolution with ϕ.)

To decide whether a potential pitch event corresponds to a real pitch event or to a spurious peak due to noise during unvoiced speech, three estimated parameters are combined into a single measure which is tested against a threshold. The current peak amplitude in the wavelet-filtered speech is compared to a running estimate of peak amplitude during loud speech, giving a relative peak amplitude parameter. The ratio of energy within the prospective pitch period in the speech signal before and after low pass filtering is the second parameter. The third parameter is a measure of the smoothness of the provisional pitch periods in the neighborhood of the pitch event under consideration. With these three parameters, a highly reliable voiced/unvoiced decision can be obtained.

This pitch detector accurately follows pitch in highly non-stationary speech.

C. Critical Bands and Speech Quality

We conducted an experimental study to determine the feasibility and quality of using the critical-band-based spectral envelope in synthesis. We considered two alternatives to complement the spectral envelope information. The first alternative is to replace the exact phase, in the frequency domain, by a simpler phase or delay information for each critical band. A preliminary study of our own, and the work of Flanagan *et al.* [1] indicates that such an approach is feasible. Such an approach results in 18 or more phase parameters for each frame, one for each critical band, which have then to be considered jointly on an acoustic subword for further transformation and representation. This representation of phase

over an acoustic segment may be quite similar to our work on the spectral envelope representation.

The second alternative is to complement the spectral envelope by the information derived from a pitch synchronous excitation. The advantage of this second approach is that the parametric description of pitch information over an acoustic segment, and thus for each frame within that segment, is extremely efficient with at most a few parameters needed for each segment.

In a simple experimental study, a binary excitation signal (using spectrally flattened glottal pulses for voiced speech and white noise for unvoiced speech) is combined with the spectral envelope to synthesize speech. The spectral envelope representation consists of bands that are of equal width and spacing in the Bark domain, and cover the frequency range from 0 to 4 kHz. The number of bands, and hence their width, is a parameter in the study. We find that some loss in quality occurs during analysis and synthesis, even with as many as 50 bands. We consider the quality of this synthesized speech to be a baseline given the particular analysis scheme and excitation signal. Reducing the number of bands to 18 results in very small, but perceptible changes. Further reduction in the number of bands is more noticeable. We adopted 18 bands as a good compromise between the achievable quality and the maximum complexity of the spectral envelope representation. Although the quality achieved by such a simple primary excitation model is not satisfactory, we consider that the experimental study provides useful information in the required number of bands.

Thus, this experiment provides experimental justification for the use of critical bands (of which there are 18 in the frequency range of interest) as an envelope representation for this analysis-synthesis approach. These results also encouraged us to seek increased representation efficiency by using the KL transformation of the spectral envelope reported in Part I.

Our method for decimating the original speech spectrum to obtain the critical band spectral envelope is described in Part I. Since the critical bands are quite narrow at low frequencies, a small problem can occur with high-pitched speakers where the critical band envelope retains some harmonic structure from the original spectrum at low frequencies. To prevent the corruption of the spectral envelope representation, a detailed envelope is generated from the original spectral magnitude by linear interpolation between adjacent harmonic peaks, and this envelope is used as the input to the decimation process.

D. Gain Normalization and Karhunen-Lóeve Transform

The first data reduction achieved by the critical band decomposition process is a 128-to-J reduction, where $J = 18$. However, because of the segmentation pre-processor, a greater reduction in the data rate is possible. Within a quasi-stationary acoustic segment (as determined by the segmentation algorithm) speech is broken down into analysis frames, which are then transformed and processed jointly. As discussed, we use the Karhunen-Lóeve Transform. In addition, since we plan eventually to quantize the critical band data, we recall that the KLT has the property of minimizing the geometric mean of the variance of the transformed vector elements, and therefore is an optimal transform for scalar quantization [11].

A useful step which limits the dynamic range of the mean vector and of the entries in the covariance matrix, consists in normalizing the critical band vectors for each frame. We let

$$p(j,n) = \frac{C(j,n)}{a(n)} \tag{1}$$

where $C(j,n)$ is the set of critical band vectors for all frames within a segment, $p(j,n)$ is the corresponding energy normalized vector set and

$$a(n) = \sqrt{\sum_{j=1}^{J} C^2(j,n)} \tag{2}$$

is the gain for each frame.

Our next step in representing the speech signal, then, is to transform the critical band vectors into the KL domain. Note that there are L vectors, one for every frame in the segment, and each vector contains J elements, one for each critical band used to decompose the spectrum. In order to transform the critical band vectors into the KL domain, we need to find the set of eigenvectors Φ of the JxJ covariance matrix Σ, which can be estimated by

$$\Sigma = \frac{1}{L}\sum_{n=1}^{L}\left[\underline{p}(n)\underline{p}^T(n)\right] - \underline{\bar{p}}\,\underline{\bar{p}}^T \tag{3}$$

Note that in contrast to the application to speech recognition, we estimated the autocovariance of the vectors, not the autocorrelation. Since the critical band vectors do not have a zero mean vector we simply subtract the estimated mean from the vectors. In order to reconstruct the speech waveform, the critical band vector mean $\underline{\bar{p}}$ is added back to the reconstructed vectors in the synthesis process.

E. KL Representation of the CB Vectors

Once the vectors are normalized and the autocovariance matrix Σ of the subword is estimated, we can solve the resulting eigenvector equation by conventional means. We can then express the CB vectors in terms of the eigenvectors, which can be considered as basis vectors since they form an orthonormal set. Thus:

$$\underline{\tilde{p}}(n) = \Phi^T\left[\underline{p}(n) - \underline{\bar{p}}\right] \tag{4}$$

F. Synthesis

The synthesis of speech is performed first by retracing the steps of the analysis to obtain a spectral envelope, then by combining this envelope with the excitation in the frequency domain and inverse transforming to obtain the synthesized speech signal.

The first step in reconstructing the spectral envelope of the acoustic segment from the KL coefficients is to take the inverse

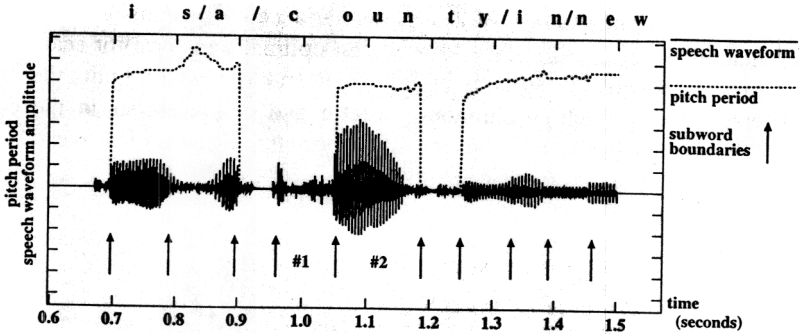


Fig. 2. Segmentation and pitch estimate.

KL transform of each vector in the segment. This can be expressed by

$$\underline{p}(n) = \Phi\underline{\tilde{p}}(n) + \underline{\bar{p}} \tag{5}$$

Since a number of coefficients may be discarded, each vector $\underline{\tilde{p}}(n)$ may be of dimension Q, $Q \leq J$, and Φ may be of dimension J x Q, where Q is the number of eigenvectors retained. The interpolation is performed by the process described in Part I. As with decimation, it is performed on a frame-by-frame basis over the segment.

The excitation signal is generated in the time domain, then transformed frame by frame to the frequency domain. To combine the spectral envelope and excitation, each frame of the excitation magnitude spectrum is multiplied by the corresponding frame of the reconstructed spectral envelope. The excitation phase is retained without modification. The resulting spectrum is inverse transformed using the overlap and add technique to obtain the synthesized speech.

G. Illustrative Examples

To illustrate this new approach to the representation of the speech spectral envelope as it applies to analysis-synthesis, we show first in Fig. 2 some of the information extracted by analysis. This figure shows the envelope of the speech waveform, the pitch period estimate profile and the segmentation information, indicated by the tickmarks on the time axis, for a portion of the sentence "Westchester is a county in New York." We observe that segments of variable duration have been extracted, and that more than one segment may occur within a single voiced speech section. We note that a short silent interval within the sentence was not selected by the segmenter for separate processing, but was grouped with an adjacent unvoiced section.

Fig. 3 shows some additional parameters extracted by analysis and the spectral envelope synthesized from this information. The two segments further processed and illustrated in Fig. 3 are segments labeled by #1 and #2 in Fig. 2.

For these two adjacent segments the spectral structures, and therefore the representations, are quite different. For each segment, we show the original critical band envelope spectrogram obtained by the decimation scheme described in Part I. From this spectral envelope representation, we extract the mean vector and the eigenvectors for each of the segments. Two eigenvectors are shown in Fig. 3. For each eigenvector, the KL coefficients are evaluated for each frame. These coefficients generally vary smoothly from frame to frame within a segment. From the mean vector, the two eigenvectors and the corresponding KL coefficients we have synthesized the critical band envelope spectrograms shown. We observe that with only two coefficients, we obtain synthesized spectrograms which already include most of the detailed structure of the originals.

H. Effectiveness of the Representation

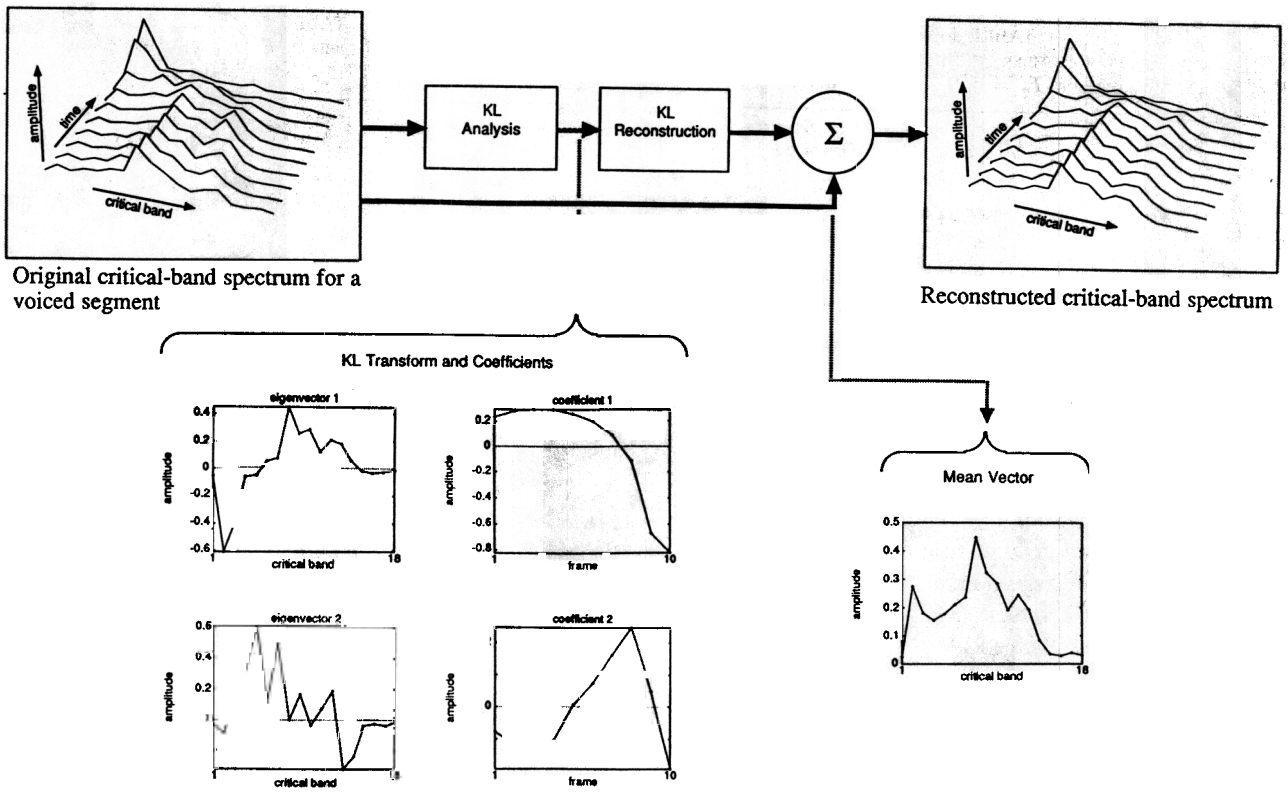
When the critical-band spectrum is decomposed using the KLT, the energy is typically concentrated heavily in the spectral mean and the low-order coefficients. The degree of concentration depends on the characteristics of the data. This degree of concentration constitutes a measure of the effectiveness of the KL representation of the critical band vector. It can be quantified by considering the eigenvalues for each segment.

Since the energy in each frame has been normalized to unity prior to the representation, the sum of the mean vector energy and of all eigenvalues is equal to unity. Thus,

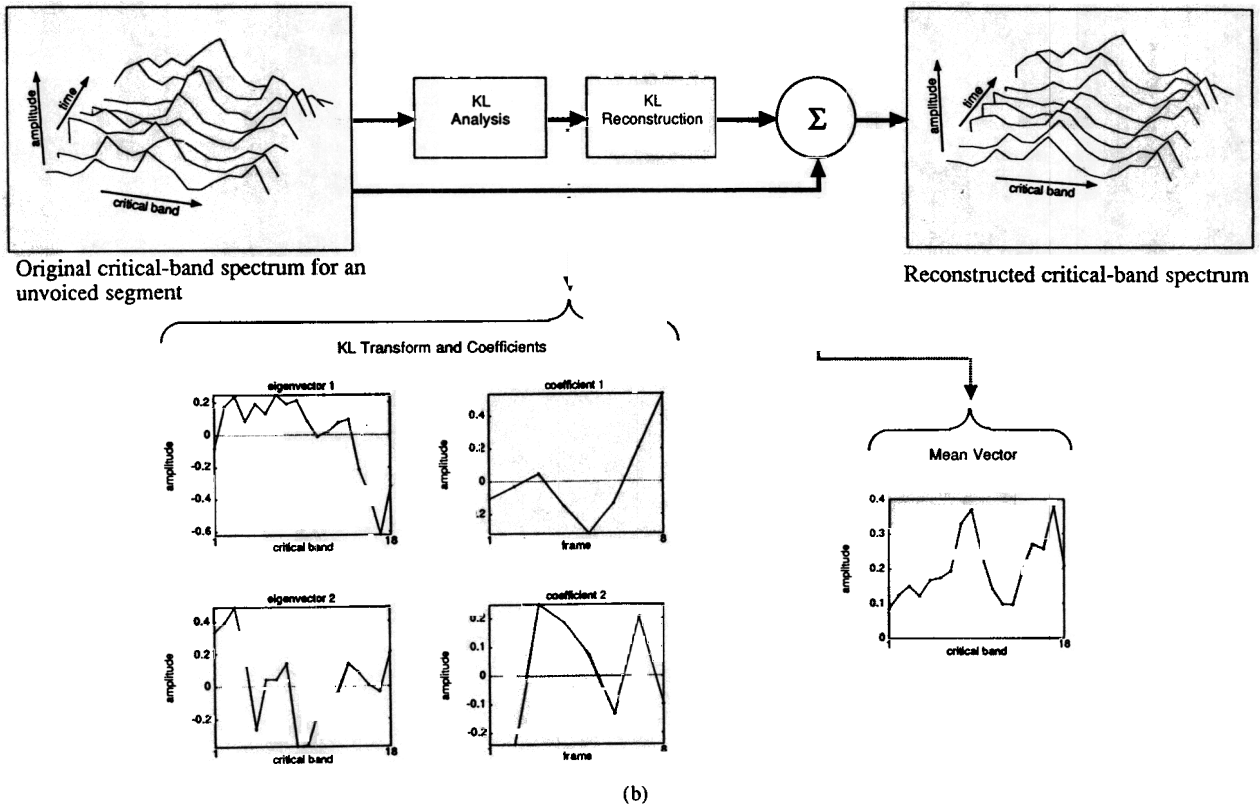
$$e_Q \qquad \sum_{j=1}^{J} \bar{p}^2(j) - \sum_{k=1}^{Q} \lambda_k \tag{6}$$

is the mean square error for a segment, where $\bar{p}(j)$ is the segmental mean for critical band j, Q is the number of retained coefficients in the representation, and $\{\lambda_k\}$ are the eigenvalues. For a subword in voiced speech, the mean and the first one or two coefficients typically account for about 99% of the energy. The same concentration of energy in unvoiced speech, where stochastic effects at the frame level are significant, can typically be found in the mean and the first six or seven KLT coefficients. In extreme cases where a vowel is highly stationary, the spectral mean can contain over 99% of the energy by itself.

Because the degree of concentration is data-dependent, it is prudent to choose the number of KL coefficients depending on the characteristics of the critical-band spectra for each subword: this allows us to achieve a consistently good approximation to the critical-band spectrum using as few bits as possible. Therefore we choose the least number of coefficients



(a)



(b)

Fig. 3. KL analysis and reconstruction. (a) Voiced segment [Fig. 2 #2]. (b) Voiced segment [Fig. 2 #1].

for each subword for which the mean square error e_Q falls below a threshold T [12]. Furthermore, since loss of the fine details of spectral envelope behavior is less perceptible in unvoiced speech than voiced speech, it is possible to use a higher threshold T_u for unvoiced subwords than the threshold T_v for voiced subwords.

TABLE I
DISTRIBUTION OF NUMBER OF COEFFICIENTS FOR EACH SUBWORD USING $T_v = 0.01$ AND $T_u = 0.07$.

No. of Coefficients	Voiced	Unvoiced
0	62	72
	484	512
	625	222
	262	11
	56	–
5	5	–

Over 96 sentences from a variety of speakers, thresholds $T_v = 0.01$ for voiced subwords and $T_u = 0.07$ for unvoiced subwords yield a distribution in the number of KL coefficients used for each subword that is summarized in Table I. For these thresholds, the average number of coefficients is 1.626. These energy threshold values are used in our experimental evaluation discussed in Section IV.

Note that decreasing the voiced threshold T_v to 0.003 raises the number of KL coefficients, as seen in Table II.

The average number of KL coefficients is now 2.260.

TABLE II
DISTRIBUTION OF NUMBER OF COEFFICIENTS FOR EACH SUBWORD USING $T_v = 0.003$ AND $T_u = 0.07$.

No. of Coefficients	Voiced	Unvoiced
0	7	72
1	134	512
2	487	222
3	490	11
4	256	–
5	95	
	19	
7	4	
8	2	

III. QUANTIZATION AND CODING

The excellent efficiency of the representation scheme for speech analysis and synthesis indicates that we may be able to reproduce good quality speech at low bit rates, but this requires that quantization and coding of the complete representation vector be included in the system. We refer to Fig. 4, a diagram of the overall speech representation and coding scheme, which shows all the components of the representation vector, as well as their approximate values after quantization. Our goal in this section is to determine and demonstrate the potential performance of our representation with quantization and coding by limiting ourselves to a simple and straightforward choice of quantizers and coders.

Summarizing briefly the results of the previous sections, we have determined that speech can be efficiently represented and synthesized from the following parameters:

1. The set of critical band representation vectors $C(j,n)$, where j is the critical band index and n is the frame number.
2. The spectral envelope segmentation information, which consists of a set of time intervals (acoustic segments) for which the speech spectral envelope is approximately constant, and a classification for each segment as voiced or unvoiced.
3. The pitch period and voicing parameter for each voiced segment.

The critical band vector $C(j,n)$, with, say $J = 18$ components per frame, can in turn be represented quite efficiently by the use of a KL transformation over the entire acoustic segment. We generate the KL transformation matrix Φ from the energy-normalized critical band vector $p(j,n) = C(j,n)/a(n)$ where $a(n)$ is the speech energy in each frame. If we know the matrix Φ, then we can find the KL representation $\tilde{p}(q,n) = \Phi^T[p(n) - \bar{p}]$. One advantage of the transformation is that, while $p(j,n)$ consists of 18-component vectors, $\tilde{p}(q,n)$ typically requires only 1 or 2 components to achieve good quality synthesized speech.

From a representation standpoint, we are trading the needed knowledge of the transformation matrix Φ for a substantial reduction in the spectral envelope information, from $p(j,n)$ to $\tilde{p}(q,n)$. Thus, referring to the diagram of Fig. 4, the speech representation is updated for every acoustic subword and consists of the eigenvectors for the subword and the critical band averages. The frame information is comprised of the KLT coefficients and the gain $a(n)$. Each of the parameters is quantized by a uniform quantizer. The only exception is the gain $a(n)$ where a logarithmic transformation is applied prior to uniform quantization. This logarithmic transformation maps the speech energy for each frame into an approximately uniform perceptual scale and allows quantization with few levels. Because the KL transformation is used on the critical band vectors, the inter-critical-band correlations have been accounted for in the representation. Thus each KL coefficient is encoded independently. We still need to account for the frame-to-frame correlation for each critical band, and thus for each of the KL coefficients.

A. Quantization Errors and Encoding Rate

We are now ready to discuss the approach to quantization and coding, and thus the digital representation of speech. We first determine an expression for the mean square error in the representation and coding of the critical band vector.

The parameters to be quantized and digitally encoded are: for each segment, $\bar{p}(j)$ which is the average value of $p(j,n)$ over all frames in the segment, and the retained vectors of the KL transformation matrix $\Phi(\phi_{j,q}, q = 1, ..., Q)$. For each frame, we quantize $a(n)$ which is the gain in the frame, and $\tilde{p}(q,n)$ which is the transformed critical band vector for the frame.

Whenever we quantize any parameter β, we shall represent it by a quantized value $\hat{\beta}$ and thus introduce an error $\epsilon = \beta - \hat{\beta}$. With this notation we can now proceed with a discussion of mean square error due to quantization and then of bit allocation for the coding scheme.

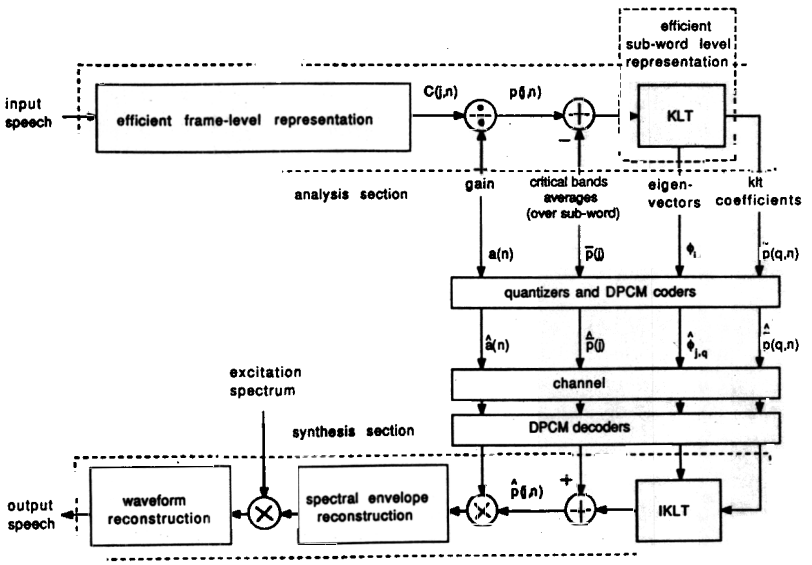


Fig. 4. Coded parameters in analysis-synthesis.

The exact critical band vector $p(j,n)$ can be expressed as

$$p(j,n) = \sum_{q=1}^{Q} \phi_{j,q}\tilde{p}(q,n) + \sum_{q=Q+1}^{J} \phi_{j,q}\tilde{p}(q,n) + \bar{p}(j) \quad (7)$$

where Q is the number of KL coefficients kept for reconstruction. The critical band vector after quantization errors are introduced can be expressed as

$$\hat{p}(j,n) = \sum_{q=1}^{Q} \hat{\phi}_{j,q}\,\hat{\tilde{p}}(q,n) + \hat{\bar{p}}(j) \quad (8)$$

whether we have discarded the $J-Q$ coefficients $\hat{\tilde{p}}(Q+1,n)$ through $\hat{\tilde{p}}(J,n)$. We now consider the errors for a given frame, and thus discard the index n for simplicity. If we define the quantized parameters by

$$\begin{aligned} \hat{\tilde{p}}(q) &= \tilde{p}(q) + \varepsilon_{\tilde{p}}(q) \\ \hat{\bar{p}}(j) &= \bar{p}(j) + \varepsilon_{\bar{p}}(j) \\ \hat{\phi}_{j,q} &= \phi_{j,q} + \varepsilon_{\phi}(j,q) \end{aligned} \quad (9)$$

where we model errors on all quantized variables as additive, uncorrelated noise with zero mean value, then we find that the expected value of the squared error between $p(j)$ and $\hat{p}(j)$ is

$$\begin{aligned} E\{\epsilon_j^2\} &= E\{[\hat{p}(j) - p(j)]^2\} \\ E\{\epsilon_j^2\} &= E\Big\{\Big[\varepsilon_{\bar{p}}(j) + \sum_{q=1}^{Q} \varepsilon_{\phi}(j,q)\tilde{p}(q) + \sum_{q=1}^{Q} \phi_{j,q}\varepsilon_{\tilde{p}}(q) \\ &\quad + \sum_{q=1}^{Q} \varepsilon_{\phi}(j,q)\varepsilon_{\tilde{p}}(q) - \sum_{q=Q+1}^{J} \phi_{j,q}\tilde{p}(q)\Big]^2\Big\}. \end{aligned} \quad (10)$$

If we ignore the cross terms of this equation and the last two summations (since the $\tilde{p}(q)$ are very small for $q > Q$, and since the product of two error terms is also small), we have

$$E\{\varepsilon_j^2\} = E\left\{\varepsilon_{\bar{p}}{}^2(j) + \sum_{q=1}^{Q} \varepsilon_{\phi}{}^2(j,q)\tilde{p}^2(q) + \sum_{q=1}^{Q} \phi_{j,q}{}^2\varepsilon_{\tilde{p}}{}^2(q)\right\} \quad (11)$$

Clearly, as we increase the quantization errors (ε) and decrease the number of retained coefficients (Q), the corresponding number of bits in the digital representation will also decrease.

Digital Rate of the Quantized Representation: We shall make explicit the relation between quantization errors and digital rates for each quantized parameter under simplifying assumptions. Our goal is to develop an approximate relation between parameters under our control, which are the number of retained KL coefficients and the quantization step size for each parameter, so as to minimize the digital rate for a given mean square representation error. These relations guide and are complemented by a simulation study carried out on examples.

If we use quantization steps Δ for each parameter, and the range of the parameter is -1 to $+1$, then for each quantized parameter the mean square error is approximately

$$\sigma_\epsilon^2 \cong \frac{\Delta^2}{12} < \frac{\Delta^2}{4} \quad (12)$$

and the corresponding rate in a direct binary representation with no coding is

$$N = \log_2 \frac{\hat{}}{\Delta} + \quad (13)$$

or

$$\Delta = 2^{1-N}$$

Therefore, if the variables $\bar{p}, \phi$ and $\tilde{p}$ have ranges $R_{\bar{p}}, R_{\phi}$ and $R_{\tilde{p}}$ respectively, we evaluate the upper bound of $\varepsilon_j{}^2$ by bounding every element in (11):

$$\begin{aligned} \varepsilon_{\bar{p}}{}^2(j) &\le \frac{\Delta_{\bar{p}}{}^2}{4} = R_{\bar{p}}{}^2 2^{-2N_{\bar{p}}} \\ \varepsilon_{\phi}{}^2(j) &\le \frac{\Delta_{\phi}{}^2}{4} = R_{\phi}{}^2 2^{-2N_{\phi}} \\ \varepsilon_{\tilde{p}}{}^2(j) &\le \frac{\Delta_{\tilde{p}}^2}{4} = R_{\tilde{p}}{}^2 2^{-2N_{\tilde{p}}} \end{aligned} \quad (15)$$

where N_x is the number of bits used to encode element x

Recall that $p(j)$ is a component of a vector with total energy unity. Therefore $\tilde{p}^2(j) \leq 1$. Similarly the $\phi_{j,q}$ are components of unit energy vectors and also bounded by one. Therefore,

$$\sum_{q=1}^{Q} \phi_{j,q}^2 \quad b_\phi \leq Q$$
$$\sum_{q=1}^{Q} \tilde{p}^2(q) \quad b_{\tilde{p}} \leq Q. \tag{16}$$

We can now express the upper bound for the mean square error in any frame by

$$\varepsilon_j^2 \leq R_{\tilde{p}}{}^2 2^{-2N_{\tilde{p}}} + R_\phi{}^2 2^{-2N_\phi} b_\phi + R_{\tilde{p}}{}^2 2^{-2N_{\tilde{p}}} b_{\tilde{p}} \qquad (1$$

Note that this bound is now independent of j. We shall use this bound as an approximate mean square error that we wish to constrain, while we minimize the number of bits required for encoding. If we define B to be the total number of bits required for each subword segment, then

$$B = JN_{\bar{p}} + LQN_{\tilde{p}} + JQN_\phi \tag{18}$$

where J is the total number of critical band filters used, L is the number of frames in the segment, and Q is the number of KL coefficients and eigenvectors transmitted. The first term in (18) corresponds to the number of bits required to encode the mean vector $\bar{p}$, the second term corresponds to the number of bits required to encode the KLT coefficients vectors $\tilde{p}$, and the third term corresponds to the bits required to encode the eigenvectors ϕ. In order to minimize B while bounding the mean square error of (17), we use Lagrange multipliers, and minimize the expression

$$B + \lambda\varepsilon_j^2 \qquad 19)$$
$$JN_{\bar{p}} + LQN_{\tilde{p}} + JQN_\phi$$
$$+ \lambda\left[R_{\tilde{p}}^2\, 2^{-2N_{\tilde{p}}} + R_\phi{}^2 2^{-2N_\phi} b_\phi + R_{\tilde{p}}{}^2 2 \right. \qquad (20)$$

with respect to $N_{\bar{p}}, N_{\tilde{p}}$, and N_ϕ. Setting the resulting derivatives to equal zero yields

$$\log_2 \quad \frac{2\lambda R_{\tilde{p}}^2 \ln 2}{} \qquad 21$$

$$2^{\log_2} \quad \frac{2\lambda R_{\tilde{p}}^2\, b_{\tilde{p}} \ln 2}{LQ}$$

$$\log_2 \quad \frac{2\lambda R_\phi^2\, b_\phi \ln 2}{} \qquad 23)$$

Now, if we set $b_{\tilde{p}} = b_\phi = Q$, then we can relate the numbe of steps in each quantizer as follows:

$$N_{\bar{p}} \quad 2^{\log_2} \left[\frac{JR_{\tilde{p}}^2}{LR_{\tilde{p}}{}^2} \right. \qquad (24)$$

$$N_{\tilde{p}} \quad N_{\bar{p}} + \quad \log_2 \quad \frac{JR_{\tilde{p}}^2}{LR_{\tilde{p}}^2} \qquad 25)$$

and N_ϕ can be expressed as

$$N_\phi \quad N_{\bar{p}} + \frac{1}{2}\log_2 \left| \frac{}{R_{\tilde{p}}^2} \right. \qquad (26)$$

Therefore, we simply need to find an adequate value for $N_{\bar{p}}$, and the optimum number of steps for the other quantizers are defined. These relations based on simple rate bounds provide the number of quantization steps (and thus the number of bits) for each parameter, as a function of a single parameter $N_{\bar{p}}$. They provide a convenient initial set of conditions for a computer simulation study.

IV. Experimental Study

A preliminary study of quality and rate by this new analysis-synthesis and coding approach has been conducted, and is described here.

A. Experimental Simulation Conditions

The parametric study was performed using the rate-mean square error relations derived above and shown in (25) and (26). Note that the ranges of the parameters to be quantized have not yet been determined in the experimental study. Quantized values for the eigenvectors and KL coefficients were encoded using first-order DPCM, with prediction coefficient of 0.9, in order to take advantage of the correlation between adjacent data samples for each parameter. By the use of DPCM, we decrease the range and thus the quantization errors for a given number of quantization levels. Simulations were performed using 5 bits to encode the gain, which changes at every frame. The primary goal of our tests was to determine the best quantization range for the KL coefficients $\tilde{p}(q,n)$, the transformation matrix components $\phi_{j,q}$, and the average critical band vector $\bar{p}(j)$. We determined the minimum range for quantization decreases the magnitude of the granular noise in the output. In order to find these ranges, we quantized the parameters using enough bits per parameter such that granular noise was negligible. We then decreased the quantization range of each parameter, one at a time, and determined the range at which overload noise became noticable. These ranges are: 0.7 for the average critical band vector, 0.7 for the KL coefficients, and 1.0 for the transformation matrix components. We also verified that 5 bits was sufficient for encoding the gain without introducing an additional noticable distortion. After finding the best ranges for the three parameters, we tested the relations between the quantization steps in the bit allocation expressions of (25) and (26). We found that the predicted number of quantization levels for each parameter was a good compromise between quality and rate.

Encoding of the Segmentation: The segmentation is performed on a frame basis and is therefore labeled by an integer with a limited range. Thus, we encode the number of frames between successive segments using a 5-bit codeword.

Encoding of Pitch: The pitch frequency is approximated by a simple piecewise linear spline with knots at every three frames. The initial pitch frequency for a voiced segment is represented by an 8-bit integer and successive knots on the

pitch profile are represented differentially using 4 bits and a sign bit. These simple codes for the segmentation and pitch result in small contributions to the overall bit rate.

In our simulations, the glottal pulse excitation is actually inserted in the frequency domain to generate a detailed speech spectrum. Because the total spectral envelope already incorporates the glottal pulse envelope, an equalization of the glottal pulse envelope is performed prior to multiplication by th global spectral envelope. Thus, the glottal pulse will only affect the detailed spectral shape and the phase of the fine speech spectrum.

B. Experimental Evaluation of Perceptual Quality and Rates

An experimental evaluation of the average bit rate of this new analysis/synthesis scheme and of the corresponding quality of the synthesized speech has been carried out. Recall that a variable number of KL coefficients is used for the spectral envelope representation based on the energy error threshold in an acoustic subword.

For the 96 sentences discussed in Section 2.8, with $N_{\bar{p}} = 5$ bits, voiced energy threshold $T_v = 0.01$ and unvoiced energy threshold $T_u = 0.07$, the average bit rate is 3.45 kb/s. Lowering the voiced energy threshold to $T_v = 0.003$ raises the average bit rate to 4.24 kb/s, and gives only a slight improvement in quality.

To gauge the speech quality of this coding scheme, we perfomed a subjective evaluation comparing the quality of this scheme at 3.45 kb/s to the DOD Proposed Standard 1016, CELP 4.8. The subjective evaluation was performed by seven subjects. A total of 30 sentences from the TIMIT database were used in the test, from 30 speakers, approximately half male and half female. Subjects heard three versions each of ten sentences: for each sentence, two types of coded speech separated by the original speech were presented, and subjects were asked to indicate which they found to be closer to the original. The order of presentation, CELP first or KL first, was randomized. Rating was on the 5-point scale of Table III, where "A" indicates the first coded sentence presented, and "B" indicates the second coded sentence.

CELP sentences were perceived by the subjects to be slightly closer to the original sentences than the KL sentences. If CELP is "A" and the KL coder is "B" on the scale above, the average rating was 2.3, and thus falls between statements "CELP is somewhat closer to the original" and "CELP and our coder differ approximately equally from the original." Thus, at 3.45 kb/s the KL coder results in a speech quality comparable to CELP 4.8.

In comparing the two schemes, note that only such broad comparisons are appropriate, since the average bit rate for the KL coder is significantly lower.

TABLE III
RATING SCHEME FOR SUBJECTIVE COMPARISON OF TWO CODERS

1	A is much closer to the original.
2	A is somewhat closer to the original.
3	A and B differ approximately from the original
4	B is somewhat closer to the original.
5	B is much closer to the original.

V. DISCUSSION AND CONCLUSIONS

The quality and low average bit rate we have achieved with our algorithm clearly indicates the merit of this speech representation and coding method. Using this new representation, we have been able to achieve completely intelligible speech at average bit rates of 3.45 kb/s for sentences. We have also verified experimentally that the critical bands of hearing provide a basis for decimating the short time speech spectrum, and is a valid and useful approach for analysis/synthesis of speech. In addition, we have shown that the use of the KLT in our representation and DPCM for coding the parameters are effective methods which provide good results. We expect that additional improvements on the quality-performance tradeoff can still be achieved. We shall mention only a few issues under study.

With respect to speech quality, we plan to refine the adaptive determination of the number of KLT coefficients used to represent each subword in the speech signal. Currently, the number of KLT coefficients retained is based on energy and is held constant during each subword. Further adaption would allow the quantization as well as number of coefficients to vary according to the particular characteristics of each segment. The use of KLT in such an adaptive scheme has a good theoretical foundation and has been applied successfully to a waveform coder [12], [13]. A second improvement of quality is possible by the refinement of the mixed excitation source so as to incorporate knowledge of the dynamics of speech into the excitation during spectral transitions.

On the quantization and coding aspects of the approach, refinements and simplifications are possible. We first observe that a substantial fraction of the total bit rate is devoted to the encoding of the KL transformation itself. This partial rate depends, of course, on the number of subword segments generated per second. As tested, the algorithm sometimes generates more subword segments than appear necessary. This over-segmentation probably does not improve quality, but is detrimental to the bit rate and needs to be refined. Another rate reduction scheme which may have merit is in the application of vector quantization for some of the parameter sets [14]. Note however that the KLT has already decorrelated the critical band vector for a whole subword segment. Thus one of the significant factors in the advantage of vector quantization has already been exploited [15].

Finally, the systematic capture of the dynamics of speech spectra available in our algorithm has also promise for an effective new approach to the synthesis of good quality speech from text.

REFERENCES

[1] J. L. Flanagan and S. W. Christensen, "Computer studies on parametric coding of speech spectra,"*J. Acoust. Soc. Am*, vol. 68, pp. 420–430, Aug. 1980.
[2] B. S. Atal, "Efficient coding of LPC parameters by temporal decomposition," in *Proc. ICASSP*, pp. 81–84, 1983.
[3] D. W. Griffin and J. X. Lim, "Multiband excitation vocoder," *IEEE Trans. ASSP*, vol. ASSP-36, pp. 1223–1235, Aug 1988.

[4] G. R. Doddington and T. B. Schalk, "Speech recognition: Turning theory to practice," *IEEE Spectrum*, pp. 26–32, Sept. 1981.
[5] W. T. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "TIMIT: The DARPA speech recognition research database: Specifications and status," in *Proc. DARPA Speech Recognition Workshop*, pp. 93–99, Feb. 1986.
[6] A. V. McCree and T. P. Barnwell, "A new mixed excitation LPC vocoder," in *Proc. ICASSP*, pp. 593–596, 1991.
[7] V. R. Algazi, D. Irvine, C. Caldwell, M. Ready, K. Brown, and S. Chung, "Speech coding by the efficient transformation of the spectral envelope of subwords, in *Proc. ICASSP*, 1992.
[8] S. Kadambe and G. F. Boudreaux-Bartels, "A pitch detector based on event detection using the dyadic wavelet transform," in *Proc. ICLSP*, Kobe Japan, Nov. 1990.
[9] S. Kadambe and G. F. Boudreaux-Bartels, "A comparison of a wavelet transform event detection pitch detector with classical pitch detectors," in *Proc. 24th Asilomar Conf. on Signals, Systems, and Computers*, Pacific Grove, CA, Nov. 1990.
[10] S. Kadambe and G.F. Boudreaux-Bartels, "A comparison of wavelet functions for pitch detection of speech signals," in *Proc. ICASSP*, pp. 449–452, 1991.
[11] R. Zelinski and P. Noll, "Adaptive transform coding of speech signals," *IEEE Trans. ASSP*, vol. ASSP-25, pp. 299–309, 1977.
[12] V. R. Algazi and D. J. Sakrison, "Encoding of a counting rate source with orthogonal functions," in *Proc. Brooklyn Polytech. Inst. Symp. on Computer Processing in Communications*, pp. 85–100, New York, 1969.
[13] V. R. Algazi and D. J. Sakrison, "On the optimality of the Karhunen-Lòeve expansion," *IEEE Trans. Information Theory*, vol. IT-15, pp. 319–321, 1969.
[14] A. Buzo, A. H. Gray, R. M. Gray, and J. D. Markel, "Speech coding based upon vector quantization advantage," *IEEE Trans. ASSP*, vol. ASSP-28. pp. 562–374, Oct 1980.
[15] T. Lookabaugh and R. M. Gray, "High resolution quantization theory and the vector quantization advantage," *IEEE Trans. Information Theory*, vol. IT-35, pp. 1020–1033, 1989.

For photographs and biographies of the authors, please see pages 194–195 of the April 1993 issue of this TRANSACTIONS.