DIXIEME COLLOQUE SUR LE TRAITEMENT DU
SIGNAL ET SES APPLICATIONS

NICE du 20 au 24 MAI 1985

KL TRANSFORM OF SEGMENTED SPEECH FOR ENHANCEMENT AND RECOGNITION

V. R. ALGAZI, M. J. READY and K. L. BROWN

Department of Electrical and Computer Engineering Signal and Image Processing Laboratory University of California, Davis Davis, CA 95616

RESUME

RESUME

La communication vocale avec l'ordinateur constitue un important développement pour accélérer l'introduction de la technologie électronique dans le monde du travail. L' usage intensif de cette technologie nécessite des techniques de traitement de la voix qui aient de bonnes performances en présence de bruit. Beaucoup de techniques de traitement de la voix ont de bonnes performances en environnements calmes, mais ces performances se dégradent en présence de bruit. Ce projet de recherche examine l'utilisation d' Algorithmes des Transformées pour traiter la parole en présence de bruit, en mettant à profit les caractéristiques des signaux vocaux.

SUMMARY

ABSTRACT

Interaction with computers by voice is an important development in accelerating the incorporation of electronic high technology into the workplace. Extended usage of this technology requires speech processing techniques that perform well in the presence of noise. While many speech processing techniques perform well in quiet environments, their performance degrades in the presence of noise. This research project investigates the use of Transform Algorithms for the processing of noisy speech that takes advantage of linguistic knowledge available from the signal.

I. Introduction

Speech signals, although generally considered to be non stationary, have a relatively well defined structure, related to linguistic events, that can be exploited in the development of speech processing tasks. Commonly, speech processing systems do not enlist knowledge of the signal characteristics in extracting information about the speech signal and, consequently, represent "ignorance models" of the significant aspects of the incoming signal. On the other hand, linguistically based systems include explicit knowledge of speech signal characteristics derived from experiments in acoustic phonetics, psychoacoustics and spectrogram reading, and, consequently, represent "knowledge based models." We propose a more natural framework for speech processing, contrasting sharply with current approaches, that exploits linguistic knowledge embodied in the signal characteristics.

II. Speech Signal Characteristics

It is generally accepted that different speech sounds are characterized by different acoustic patterns. Fricatives, for example, are characterized by their noise-like time structure and are highpass in nature; while vowels are quasi periodic and have a distinctly different formant structure. The corresponding spectral pattern, which exhibit stationary behavior for the duration of each basic speech sound, serve as phonetic event indicators. Spectrogram readers, such as Victor Zue, demonstrate that the acoustic patterns visible from broadband spectrograms encode sufficient information to identify phonetic events. In fact, Ronald Cole asserts that spectral information is sufficient for identification of phonetic segments with a first choice accuracy of 85%[1]. Concurrently, speech synthesized from LPC models substantiate broadband representations.

We exploit this linguistic knowledge by segmenting the speech signal into linguistically contrasting parts and, in turn, transforming the segments into the spectral domain. This leads to a signal processing model that considers the speech to be composed of piecewise stationary segments.

This approach has several beneficial implications. First, it provides a natural, linguistic segmentation of the speech based on signal characteristics. Typically, other systems arbitrarily partition the speech into uniform analysis frames (typically 10msec), reflecting the non stationary assumption, not synchronized to the speech characteristics. Secondly, the segments can be processed as single, stationary events exploiting correlations within the segment. In other systems, the analysis frames are processed independently and ignore interframe correlations. Thirdly, processing the whole segment jointly results in a more robust extraction of the speech characteristics in noise because the duration of the segments are generally longer than the short analysis frame used by other systems[2].

For specific speech processing applications, the new approach offers important advantages. In recognition systems it reduces the possible reference templates to utterances having the same segmental decomposition; excludes irrelevant data that may cause confusion between similar, though distinctly different sounds, e.g. B vs. D.

Within the context of speech enhancement and compression, the new approach provides a more stable estimate of speech signal characteristics (especially important in noisy environments); allows for a fixed processing strategy for the entire segment rather than a new strategy for each time frame independently; avoids joint processing of adjacent portions of the signal with radically different characteristics (such as voiced to unvoiced regions) that can occur when using an analysis frame not synchronized to the signal characteristics.

III. Karhunen Loeve Transform (KLT)

Conventionally, the KLT has been used as a tool for dimensionality reduction and data compression. It achieves dimensionality reduction by capturing patterns of correlations among observed variables. Because it uncovers patterns of intercorrelations, it can be used as a mathematical technique to analyze relationships and interdependencies in a set of quantitative variables and to reveal underlying structural patterns. The KLT reduces a large set of correlated variables to a set of ranked uncorrelated latent variables, indicating the relative dominance of each of the latent patterns within the data. We apply the KLT to the speech spectral data decomposed into adjacent bands to characterize each segment independent of its duration. It is also well known that the KLT provides optimum mean squared filtering of each segment in the presence of white noise. This leads to processing techniques robust in noise.

Exploitation of this characterization depends on the end application. We give two examples below.

IV. Applications

We demonstrate the new approach via two important speech processing tasks: speaker dependent, isolated word recognition and enhancement of speech degraded by additive white noise. In these examples, we have manually segmented the speech into voiced, unvoiced and silence regions via graphical display ignoring any nonstationarities within the regions.

A. Recognition [see 3.]

Table 1 summarizes the spectral patterns in 16 adjacent bands captured by different KL components (indicated by #) for two repetitions of each of the fricatives and stop consonants. The table, for example, indicates that component 1 for each of the repetitions of the unvoiced fricative /s/ (S1 and S2) is a high frequency average; likewise, it indicates that component 2 for S1 and S2 is a comparison of high frequency bands.

PHON	AVERAGES								COMPARISONS							
EME	l	m	m+	h	l/m+	m+	l/l	l/h	m/m	l/m+	m/m+	m/m+	m+/m+	m+/h	l/h	h/h
S1 (UV)				1												2
S2 (UV)				1												2
SH1(UV)					1							2				
TH2(UV)					1							2				
ZH1 (V)			1							2,3		2				
ZH2 (V)			1							2,3		2				
K1 (UV)		1														
K2 (UV)		1														
T1 (UV)	1															2
T2 (UV)	1															2
D1 (V)	1								2,3,4						3,4	
D2 (V)	1								2,3,4						3,4	
Z1 (V)	1					2									2	
Z2 (V)	1					2									2	
G1 (V)	1	2														
G2 (V)	1	2														
V1 (V)	1,2	4				2										
V2 (V)	1,2	4				2										
TH1(UV)	1					2,3,4,5	2,3		5,6							
TH2(UV)	1					1,2,4,5	2,3		4,5,6							
F1 (UV)	1					2,3,3	3,4	6,7,8		9	10					
F2 (UV)	1					2,3	3,4	6,7,8		9	10					
B1 (V)	1					2,4,3	3,4	6,7,8							9,10	
B2 (V)	1					2,3	3,4	5,6,7							9,10	
P1 (UV)	1					3,4	5,6	4,5,7								8,9
P2 (UV)	1					2,3	3,4,7	5,6,7								8,9
TH1 (V)	1,2					3,4										
TH2 (V)	1,2					3,4										

Table 1. Patterns extracted by indicated components for 2 repetitions of the frocatives and stop consonants where l=freq bands 1-4, m=freq. bands 5-8, m+=freq. bands 9-12, h=freq. bands 13-16; UV=unvoiced and V=Voiced.

The KL method, thus, represents a hierarchical pattern extraction scheme. The first component, accounting for the largest proportion of the variance, captures the generalized pattern of interdependency among all the bands (e.g. concentration of high-frequency energy). The second component, explaining a smaller proportion of the variance, begins to uncover more subtle patterns of interdependencies (e.g. the interdependent structure of those high frequency bands specified by component 1). Higher order components, in turn, will uncover even more subtle patterns of interrelationships.

In some instances, the first component by itself establishes a differentiating pattern, one which distinguishes its phoneme from all other phonemes. For example, /s/, /sh/, /zh/, and /k/ can all be identified by their first component. In contrast, the /f/, /p/, /b/ group of phonemes requires nine to ten components to uncover their differentiating characteristics.

Because the KL-transform captures spectral relationships and interdependencies, actual frame number does not greatly influence it, and, consequently, no complicated time alignment scheme is necessary. For example, the first two KL components for two repetitions of the phoneme /s/, one with half the time duration of the other, are illustrated in table 2. Since patterns of interdependencies do not change substantially with duration, the component coefficients likewise, show little change.

DURATION (FRAMES)	COMPONENT	FREQUENCY BAND 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
20	1	-.009	-.002	-.007	-.01	-.005	-.004	-.01	-.07	-.01	-.05	-.1	-.1	-.6	-.6	-.3	-.3
	2	-.03	.05	.03	.02	.01	.002	.01	.06	.02	.04	.1	.1	.5	-.08	-.2	-.7
10	1	-.03	-.01	-.007	-.01	-.003	-.005	-.01	-.06	-.01	-.03	-.1	-.1	-.6	-.6	-.3	-.5
	2	-.003	.1	.04	.02	.01	.009	.01	.02	.03	.07	.1	.2	.5	-.08	-.2	-.7

Table 4: Illustration of the inherent time alignment performed by the KL Transform on the phoneme /s/.

Table 2. First two spectral patterns for /s/ for different durations.

Table 1 illustrates that different repetitions of the same phoneme are characterized, as expected, by similar spectral patterns. A simple template matching of component coefficients, therefore, serves as a classification strategy. A flow diagram of the algorithm is given in figure 1. The internal pattern extraction stage, implemented for each segment, consists of the following: (1) calculation of a covariance matrix from the spectral data; (2) calculation of the set of eigenvectors and eigenvalues for the covariance matrix determined in (1).

After each segment acquires a set of eigenvectors and eigenvalues, the algorithm pursues divergent paths. In training mode, the eigenvectors and eigenvalues, along with the segmental decomposition, are simply stored in reference memory; in recognition mode, each segment enters a classification stage, which consists of two steps: (1) Comparison of the segmental decomposition of the input word with corresponding sequences of the template words. If a match occurs with only one template, classification is considered complete at this step; otherwise, a reduced reference library, consisting only of those templates that achieved a match, is formed for use in step (2). (2) Template matching stage in which distances between the unknown input and the allowed

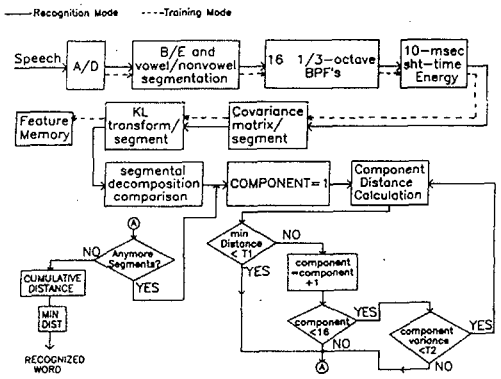


Figure 1. Flow diagram of speech recognition algorithm.

reference components are computed, distances between components (i.e. between coefficient loadings) being calculated with the Euclidean distance metric. The template matching stage is a threshold-controlled process, which compares components one by one. When a sufficient criterion is met, the distance calculation is halted and the segment is classified. Threshold 1(T1), for example, performs an inter-distance comparison. Thus, if the segment's distance from one of the templates is substantially smaller than from the others, comparison of components stops. If, however, the distances are clustered around some common value, the Euclidean distance between the next higher order component and its corresponding component in each of the allowed templates is computed. This distance is summed with any previous component distances and subjected to the T1 test again. Threshold 2 (T2) assures that each component contains valid speech information. Thus, if a component's variance (indicated by its corresponding eigenvalue) approaches that of the background noise variance (calculated during nonspeech activity), T2 halts further comparison of components. Finally, a comparison involving all sixteen components represents a forced completion condition.

Segmental distances are then cumulated to classify each word. Note that the algorithm will give equal importance to all eigenvectors, as long as the eigenvalue exceeds the noise variance.

The algorithm was tested using two different vocabularies: the digits and the confusable E set. In each case, one repetition was selected as the training set; the other repetitions were used as the test set. The algorithm achieved a 100% recognition rate on ten repetitions of the digit vocabulary. Although only a superficial analysis of the E set was conducted with three repetitions, the algorithm performed successfully--100% recognition--and made use of up to 3 eigenvectors to achieve recognition.

The algorithm was also tested in a background, computer-generated white noise. For comparison, a benchmark algorithm, which uses a filter bank for preprocessing and the Sakoe/Chiba DP algorithm for time alignment [4], was also implemented and evaluated in white noise. The results are indicated in figure 2. The new algorithm experiences only a gradual deterioration in performance even at high noise levels because the KL transform will perform an automatic noise removal on the statistically uncorrelated white noise. A modified distance measure, which weighted the KL components by their eigenvalues, was used in these noise tests.

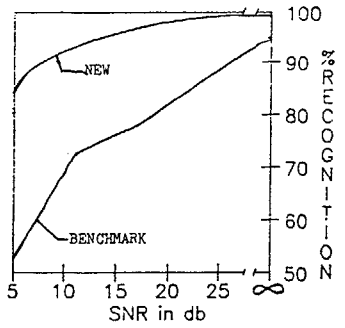


Figure 2. Performance comparison between new recognition algorithm and benchmark algorithm.

B. Enhancement of Noisy Speech [see 5.]

Speech degraded by an additive, uncorrelated background noise is annoying to listen to, fatigues listeners and reduces the overall quality. The goal of speech enhancement systems is to improve some quality aspect of the signal. We assume that only the noisy signal is available. The concept of speech quality concerns the total auditory impression of speech on a listener. It is appraised on the basis of preference, loudness, intelligibility, recognizability of properties of the original speaker's voice and other speaker characteristics. However, precise definitions of these attributes and their correlates to the acoustical signal are not known. Consequently, the approach taken here is to reduce the background noise with the hope that the important distinctive features of the speech signal are preserved.

In contrast to recognition systems which require only gross distinguishing features of the speech signal, enhancement systems require that important detail of the signal must be tracked to retain subtle perceptual cues. Our approach estimates the speech in 12.8 msec time frames (characteristic of human perception) by shaping the spectral envelope of the DFT spectral magnitude, thus tracking local behavior, but takes advantage of intra and interframe correlations within each segment. Implicitly, the signal is decomposed into a fine spectral structure and gross spectral envelope. We estimate only the envelope but retain the fine structure for resynthesis of the time domain signal. By partitioning the speech into piecewise stationary segments, classical (stationary) estimation techniques can be applied to each segment independently.

Figure 3 shows a high level block diagram of the system. The spectral envelope is estimated by filtering the local noisy speech envelope, obtained via the bank of filters (BOF) as in the recognition system, with a fixed filter derived for each segment. The filter characteristics are derived from intra and interframe noisy speech spectral characteristics, measured via the KLT, within each segment and the average noise characteristics measured during non speech activity. The optimum mean squared filter characteristics in KL space (for white noise) are:

$$a(k)=\frac{\lambda(k)-\sigma^2(k)}{\lambda(k)}$$

where $\lambda(k)$ are the eigenvalues of the spectral data covariance matrix and $\sigma^2(k)$ are the corresponding noise variances. The time domain estimate is obtained by shaping the DFT spectral magnitude of the noisy speech to have the estimated spectral envelope and inverse transforming.

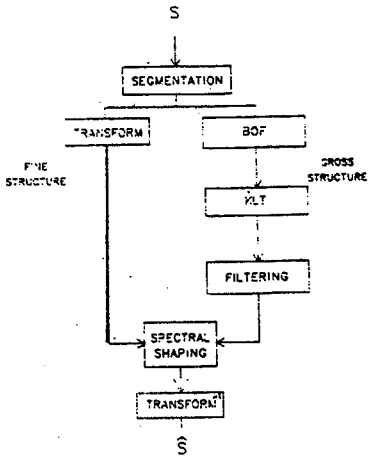
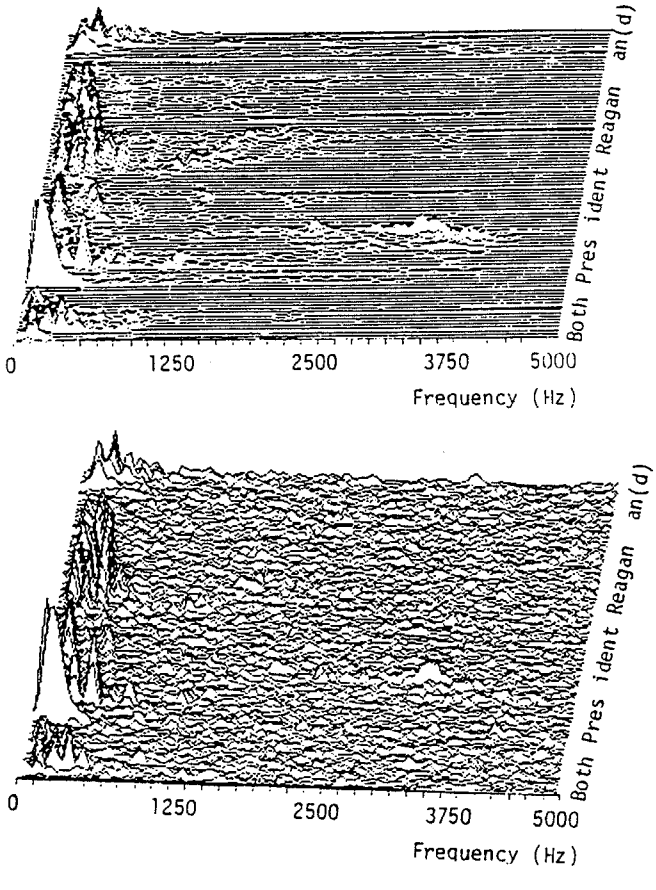


Figure 3. High level block diagram of the new speech enhancement system.

Figure 4 shows the periodogram of clean, noisy and enhanced speech for the sentence "*Both president Reagan and Democratic challenger ....*" The signal to noise ratio is about 3db (quite loud). Many of the speech features are obscured by the noise. The enhanced speech recovers the major features of the speech while substantially suppressing the noise. In informal listening tests, the speech enhanced by the new algorithm sounds better than the same speech enhanced by spectral subtraction [6] and Boll's[7] algorithm.

V. Conclusions and Future Work

The segment analysis sections of the recognition and enhancement algorithms are identical. This suggests that the framework established in this paper may apply to other speech processing tasks. Currently, we are investigating new methods,



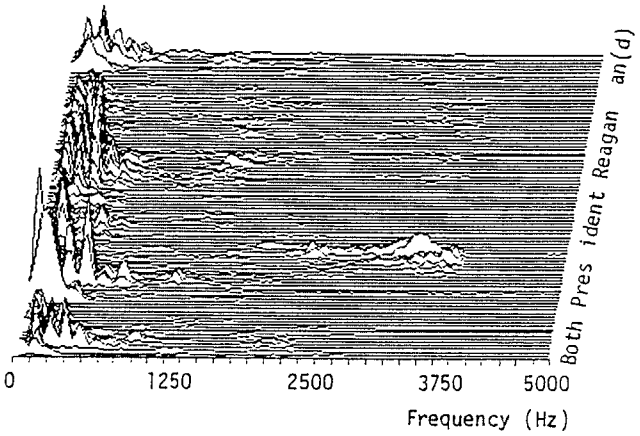


Figure 4. Spectrograms of clean, noisy and enhanced speech.

based on this work, of extracting LPC parameters from noisy speech. Note that the segmentation of speech into stationary segments was performed graphically. A detection theory based approach to segmentation is under development [8] and promises to perform quite well in noisy environments.

References

1. Cole, R. A., "Performing fine phonetic distinctions: Templates vs. features," *Conf. Toward Robustness in Speech Recognition*, Santa Barbara, Calif., Nov. 1983.
2. Crystal, H. T. and House, A. S., "Segmental durations in connected speech signals: Preliminary results," *J. Acoust. Soc. Am.*, no. 3, pp. 705-716, Sept, 1982.
3. Brown, K. L. and Algazi, V. R., "Discrete utterance speech recognition without time alignment," *IEEE ICASSP*, Tampa, Fl., Mar., 1985.
4. White, G. M. and Neely, R. B., "Speech recognition experiments with linear prediction, bandpass filtering and dynamic time warping," *IEEE Trans. Acoustics, Speech and Signal Proc.*, pp. 183-188, Apr 1976.
5. Ready, M. J., "Enhancement of Noisy Speech Based on Speech Production and Perceptual Models," *PhD thesis*, p. University of California, Davis, Ca., Dec. 1984.
6. Lim, J. S. and Oppenheim, A. V., "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586-1604, Dec. 1979.
7. Boll, S. F., "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech and Signal Proc.*, vol. ASSP-27, no. 2, pp. 113-120, Apr. 1979.
8. Parkes, S., Algazi, V. R., and M. J. Ready, *Segmentation of noisy speech using transforms techniques*, in preperation.