homomorphic deconvolution to shallow-water marine seismology —Part I: Models," *Geophysics*, vol. 39, pp. 401–416, Aug. 1974.
[5] T. G. Stockham, Jr., T. M. Cannon, and R. B. Ingebretsen, "Blind deconvolution through digital signal processing," *Proc. IEEE*, vol. 63, pp. 678–692, Apr. 1975.
[6] D. P. Skinner, "Real-time composite signal decomposition," Ph.D. dissertation, Univ. Florida, Gainesville, 1974.
[7] D. G. Childers and A. E. Durling, *Digital Filtering and Signal Processing*. St. Paul, MN: West, 1975.
[8] G. D. Bergland, "A fast Fourier transform algorithm for real valued series," *Commun. Ass. Comput. Mach.*, vol. 11, pp. 703–710, Oct. 1968.
[9] J. W. Hartwell, "A procedure for implementing fast Fourier transform on small computers," *IBM J. Res. Develop.*, vol. 15, pp. 355–363, Sept. 1971.
[10] J. Allen, "Computer architecture for signal processing," *Proc. IEEE*, vol. 63, pp. 624–633, Apr. 1975.
[11] W. H. Specker, "A class of algorithms for $\ln x$, $\exp x$, $\sin x$, $\cos x$, $\tan^{-1} x$, $\cot^{-1} x$," *IEEE Trans. Electron. Comput.* (Short Notes), vol. EC-14, pp. 85–86, Jan. 1965.

Residual Energy of Linear Prediction Applied to Vowel and Speaker Recognition

HISASHI WAKITA

Abstract—**Recognition of steady-state vowels based on the residual energy of linear prediction was ascertained to be useful for a recognition system in which the reference data are taken from the intended speaker. Sharp speaker selectivity based on a threshold criterion suggests that the use of the residual signal energy may also be useful for speaker identification, especially for speaker screening in a large population.**

Introduction

In applying the linear prediction technique to speech recognition and speaker identification, the energy of the residual signal seems to be effective as an alternative to the filter coefficients and reflection coefficients themselves. Recently it has been shown that the residual signal energy of the linear prediction filter satisfies the distance measure criterion and that its use for word recognition has been quite successful [1]. There are various methods for utilizing the residual signal energy for recognition purposes. This correspondence investigates the effectiveness of two of these methods in their application to speech recognition and speaker identification.

Residual Signal Energy

This study considers a problem of identifying stationary vowels uttered in the context of consonant–vowel–consonant. We assume reference filters for m stationary vowels $/K_j/$ $j = 1, 2, \cdots, m$. Each reference filter is a linear prediction inverse filter [2] which is optimally computed from a corresponding reference vowel. Our problem is to determine the most likely vowel for an unknown stationary vowel X. For this purpose, the following four residual energies labeled as (A), (B), (C), and (D) in Fig. 1 can be computed.

Manuscript received June 6, 1975; revised October 24, 1975. This work was supported by the Office of Naval Research under Contract N00014-67-C-0118.

The author is with the Speech Communications Research Laboratory, Inc., Santa Barbara, CA 93109.

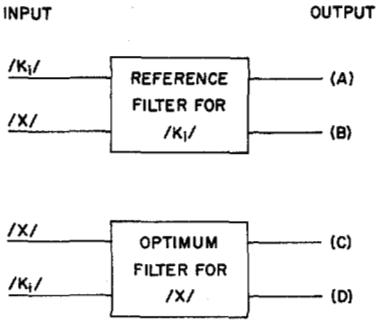

Fig. 1. Residual signal energy of linear prediction inverse filter (see text for definition of (A), (B), (C), and (D) and see Table I for their mathematical expressions).

(A) Residual energy of the reference sound $/K_j/$ when it is passed through the reference filter for $/K_j/$.
(B) Residual energy of /X/ when it is passed through the reference filter for $/K_j/$.
(C) Residual energy of /X/ when it is passed through the optimum inverse filter for /X/.
(D) Residual energy of the reference sound $/K_j/$ when it is passed through the optimum inverse filter for /X/.

For a decision based on the comparison of the above residual energies, six pairs are conceivable, i.e., 1) (A)–(C), 2) (B)–(C), 3) (A)–(D), 4) (B)–(D), 5) (A)–(B), and 6) (C)–(D). Intuitively, although it is not strictly theoretically true, 2) and 3), and 5) and 6) seem to be respectively equivalent. The comparison of residual energies by 1) and 4) does not theoretically give a minimum for a correct pair. Thus, 2) and 5) were chosen for experiment in this study. Each residual energy is defined in Table I and the decision rules for 2) and 5) are given in Table II. Method I in Table II is identical to the one used in [1].

Experiment

Nine American vowels (/i/, /ɪ/, /ɛ/, /æ/, /ʌ/, /ɑ/, /u/, /ʊ/, and /ɝ/) in the context of /hVd/, where V denotes any of the above vowels, were uttered by 16 speakers (9 males and 7 females). Ten of the speakers grew up in California and the rest were residents of California for 5 to 15 years. Thus the vowel /ɔ/ was eliminated from the experiments since /ɔ/ tends to merge into /ɑ/ among Californians. Six reference speakers (3 males and 3 females) were randomly chosen from these speakers and recognition of the above vowels based on the decision rules in Table II were performed by utilizing the reference information extracted from each of the reference speakers. A total of 288 vowels (9 vowels × 16 speakers × 2 repetitions) were used for recognition with each of the reference speakers. The vowels for this experiment were collected all in one session for both reference and test. However, the vowels used for the test do not include those used for extracting the reference information.

The words were digitized with a sampling frequency of 10 kHz and the autocorrelation method of linear prediction was applied to these speech data with a constant frame size of 30 ms, a frame shift of 6.4 ms, a +6 dB/octave preemphasis, and a Hamming window. For each frame 10 filter coefficients were computed to obtain the residual signal energy. The filter coefficients and the residual signal energy of each reference sound were averaged over three frames in the stationary portion of the sound and these were stored as reference information. The most stationary portion was determined by visually detecting the most stable portion of the first three formant frequencies.

TABLE I
MATHEMATICAL DEFINITIONS FOR RESIDUAL SIGNAL ENERGIES IN FIG. 1

A	$E_{K_j}^{K_j}$	$= \sum_{i=0}^{M} a_i^{K_j} r_i^{K_j}$
B	$E_X^{K_j}$	$= \sum_{i=0}^{M} \sum_{l=0}^{M} a_i^{K_j} a_l^{K_j} r_{\lvert i-l \rvert}^{X}$
C	E_X^X	$= \sum_{i=0}^{M} \alpha_i r_i^X$
D	$E_{K_j}^X$	$= \sum_{i=0}^{M} \sum_{l=0}^{M} \alpha_i \alpha_l r_{\lvert i-l \rvert}^{K_j}$

$a_i^{K_j}$: filter coefficients of the reference filter for $/K_j/$

α_i : filter coefficients of the optimum filter for /X/

$r_i^{K_j}$: normalized autocorrelation function of $/K_j/$

r_i^X : normalized autocorrelation function of /X/

TABLE II
DECISION RULES BASED ON A MINIMUM DISTANCE PRINCIPLE

Method		Decision Rules
I	B vs. C	$E_X^{K_i} / E_X^X$ = min over all K_j => $/X/ \in /K_i/$
II	A vs. B	$E_X^{K_i} / E_{K_i}^{K_i}$ = min over all K_j => $/X/ \in /K_i/$

RESULTS

Fig. 2 shows examples of the recognition results for two different reference speakers. It is known that the residual signal energy is quite sensitive to interspeaker differences. While the average rate of correctly identified vowels for the reference speakers was 99 percent in Method I, in Method II it was 90 percent. The average rate of correctly identified vowels for the rest of the speakers was 44 percent in Method I and 30 percent in Method II. From these results it is expected that the methods may have fairly sharp speaker selectivity if a certain threshold is set as a decision criterion instead of merely taking a minimum. Examples of the results of recognition experiments based on empirically determined thresholds are shown in Fig. 3. The speaker selectivity is seen to be quite satisfactory.

CONCLUSIONS

1) Method I gave better results than Method II in steady-state vowel recognition. Method I is also more reliable than Method II. In fact, Method I has so far been found to be more robust for long-term intraspeaker fluctuations than Method II. Method II showed an approximately 7 percent deterioration in recognition rate for the vowels recorded from two male and two female speakers six months after the reference data were extracted, whereas Method I resulted in only 2 percent deterioration. Method I may be used for a single-speaker recognition system for steady-state vowels, in which the reference information is extracted from the intended speaker. Since similar results as typified in Fig. 2 can be expected for a larger population, normalization of interspeaker differences would be needed for a multispeaker system for steady-state vowels.

2) Computationally, Method II is more efficient than Method I. In Method II, computation of the residual energy (B) in Table I requires the autocorrelation coefficients of the input signal only. In Method I, on the other hand, computation of the optimal filter for a segment of the input signal is required besides the computation of residual energy (B).

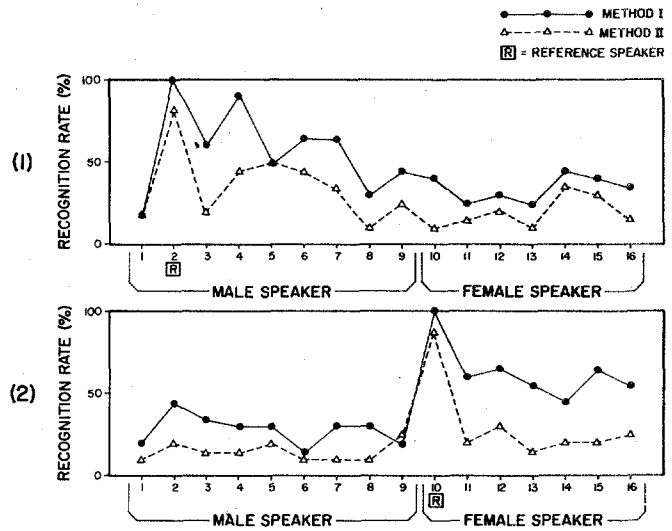

Fig. 2. Examples of stationary vowel recognition based on a minimum distance principle: (1) male reference speaker and (2) female reference speaker.

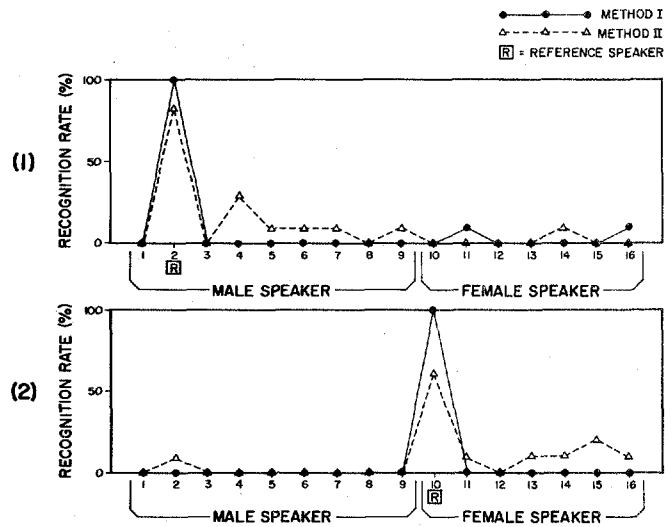

Fig. 3. Examples of stationary vowel recognition based on a threshold criterion: (1) male reference speaker and (2) female reference speaker.

3) Both methods are quite sensitive to interspeaker differences. Thus the methods may be suitable for speaker recognition based on steady-state vowels, especially for speaker screening in a large population in order to seek out a small number of the most likely candidates for further detailed examination.

REFERENCES

[1] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing (Special Issue on IEEE Symposium on Speech Recognition)*, vol. ASSP-23, pp. 67–72, Feb. 1975.

[2] J. D. Markel, "Digital inverse filtering, a new tool for formant trajectory estimation," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 129–137, June 1972.