Ann. Bull. RILP, No. 13, 137-143 (1979)

EFFECT OF ACOUSTIC FEATURE PARAMETERS OF SPEECH ON PERCEPTUAL IDENTIFICATION OF SPEECH

Shuzo Saito and Kenzo Itoh*

Introduction

There are two kinds of perceptual functions in the identification of speech; these are the identification of phonemic or semantic information and the identification of the tonal information of speech sounds. The identification of tonal information includes that of the individual speaker, of sex, of age, of emotional condition and so on. Among these, the perceptual identification of the individual speaker is perhaps the most important.

Several studies have been made on the auditory identification of a speaker, with reference to the effects of the transmission frequency band and its duration time,[1, 2] of fundamental frequency,[2] of stimulus content and its duration time[3] and so on.[4, 5, 6]

In this paper, two types of the feature parameters of speech, the frequency spectrum envelope and the excitation source signal were transformed directly by the use of the PARCOR speech analysis-synthesis technique and the effects of these feature parameters on auditory identification of a speaker were studied.

Experimental Procedure

Speech analysis-synthesis system

To manipulate the feature parameters, that is, the frequency spectrum envelope and the excitation source signal, the PARCOR speech analysis-synthesis system was used. A block diagram of the test system is shown in Fig. 1. Speech input was passed through a low pass filter of 3.4 kHz cutoff frequency and then its amplitude was digitized into 12 bits every 125 microseconds of sampling period. This digital signal was then fed to the

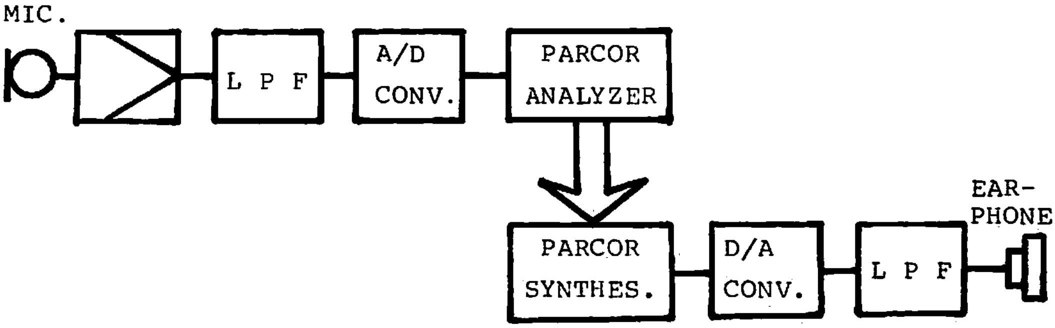

Fig. 1. Block diagram of the test system

* Musashino Electrical Communication Laboratory, N. T. &T.

PARCOR analyzer and the feature parameters of speech input were extracted in every frame period of five milliseconds. The frequency spectrum envelope was represented by the use of k parameters of up to twelve orders, and the excitation source signal was expressed in the fundamental frequency, the speech power, and the ratio of voiced and unvoiced speech powers. In the synthesis stage, these feature parameters were transformed to synthesize various kinds of test sounds used for auditory identification of the speaker. The system gain of the auditory test system was set at 0 dB in the orthotelephonic response.

Speech materials

Three kinds of test speech materials were used. (1) The first was the utterance of a sentence of about 3 seconds length. (2)The second was an utterance of the same sentence, but the timing of word spurts in the sentence was forced to imitate the utterence of a specified speaker. (3) The third was utterances of five Japanese vowels /a, i, u, e, o/, where each was extracted in 335 milliseconds lengths from utterances of the sustained vowels. Five vowels are presented successively at intervals of 67 milliseconds length.

Subjects

Speech materials were uttered by five male speakers. Ten male listeners with normal hearing were engaged in the speaker identification tests in a monaural listening situation. All listeners are colleagues of the five speakers and were familiar with their voices, thus it was easy to identify each speaker aurally.

Testing method

To evaluate the effects of feature parameters of speech on speaker identification, the analyzed outputs of the feature parameters were transformed and then fed to the synthesizer for the use of its control signals. Listeners were requested to identify the speaker for each synthesized output signal. This type of "naming" method is used in most speaker identification tests, but the ABX method is used supplementally in a few speaker identification tests.

Results

Experiment 1

The effects of the frequency spectrum envelope and the excitation source signal on speaker identification were tested using three kinds of speech materials (1), (2) and (3). The "naming" method was used for listener response. Results are shown in Fig. 2. In this figure, the abscissa represents the test conditions and the ordinate is the identification rate of the speaker. Test condition (a) was for the case of the test signal of the original speech, in other words, the input signal of the PARCOR analysis-synthesis system. Test condition (b) was for that of the PARCOR synthezer output in which whole control signals for synthesis were operated. Comparing the identification rate of condition (b) with that of condition(a), there are no significant differences for speech materials (1) and (2), but it can be seen that there is a small difference for speech material (3). It is supposed that some deterioration resulted from the analysis-synthesis technique affects only vowel stimuli, thus making it a rather low identification rate. In test condition (c), the control signal of the frequency spectrum envelope for the PARCOR synthesizer is the same as when it is analyzed, but the fundamental frequency is set to an averaged single frequency. It can be seen that there were small and same order deteriorations in the speaker identification rates for the three kinds of speech materials. Test condition (d) was the inverse of test condition (c), that is, the control signals of the excitation source signal were fed to the PARCOR synthesizer, but those of the frequency spectrum envelope were deleted. So the synthesizer output signals of the test condition (d) were the same as the excitation source signals. Comparing the identification rates of test

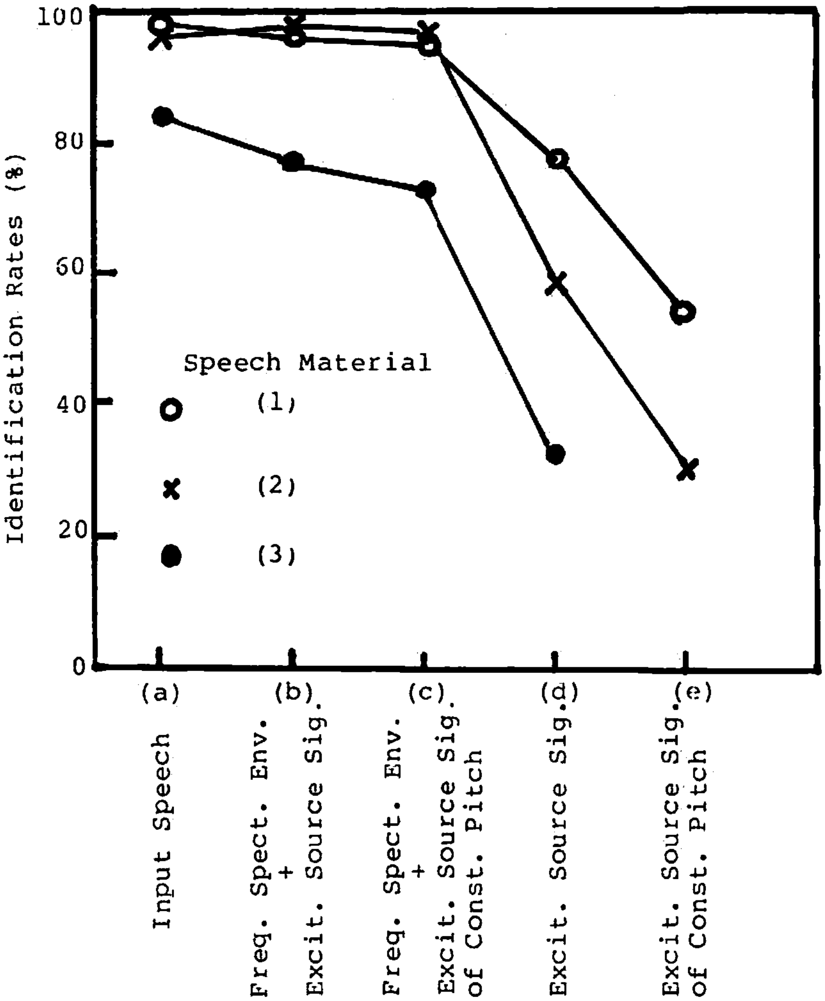

Fig. 2 Results of speaker identification test, Experiment 1.

conditions (c) and (d), it can be seen that there are not only remarkable deteriorations of identification rates in test condition(d), but also that there is a significant difference between the two speech materials, (1) and (2).

In test condition (e), the control signals of the frequency spectrum envelope are not only deleted, but also the fundamental frequency of synthesized output is set to the averaged signal value. It seems that the speaker identification rates are worst in test condition (e) and the difference in identification rates between the two speech materials (1) and (2), is comparable to that of test condition (d). The difference between speech materials (1) and (2) is in the temporal characteristics of speech as described before.

To evaluate the experimental data of the speech materials (1) and (2), variance analysis was performed. Results show that the two main factors, that is, the effects of the frequency spectrum envelope and the fundamental frequency of speech were highly significant and the effect of speech materials was also significant, but that of speaker was not significant. It may be concluded from the results of Experiment 1 shown in Fig. 2 that (1) the effect of the frequency spectrum envelope on speaker identification was prominent among the various factors, (2) the effects of the fundamental frequency and the temporal characteristics of utterance on speaker identification were significant provided that the frequency spectrum envelope of speech was removed from the speech signal.

Experiment 2

In previous experiment (Experiment 1), the effect of the frequency spectrum envelope on speaker identification was tested for two test conditions, in which the numbers of the k parameters in PARCOR analysis were 12 and 0. In this experiment, the numbers of the k parameters were varied as 12, 7, 4 and 2 to evaluate the effect of the frequency spectrum envelope more precisely.

As it was found in Experiment 1 that the effect of the temporal characteristics of utterance were not significant in both test conditions (b) and (c), speech material (2) was deleted in this experiment. The test procedures were the same as in Experiment 1 (c), except for the test conditions described above.

The results are shown in Fig. 3. In this figure, the abscissa represents the number of k parameters and the ordinate is the identification rate of speaker. It can be seen that the effect of the number of k parameters was considerable for speech material (1), but was less for speech material (3). It may be concluded that the reduction of the number of k parameters affects the dynamic characteristics of speech and thus considerable deterioration of speaker identification is brought about.

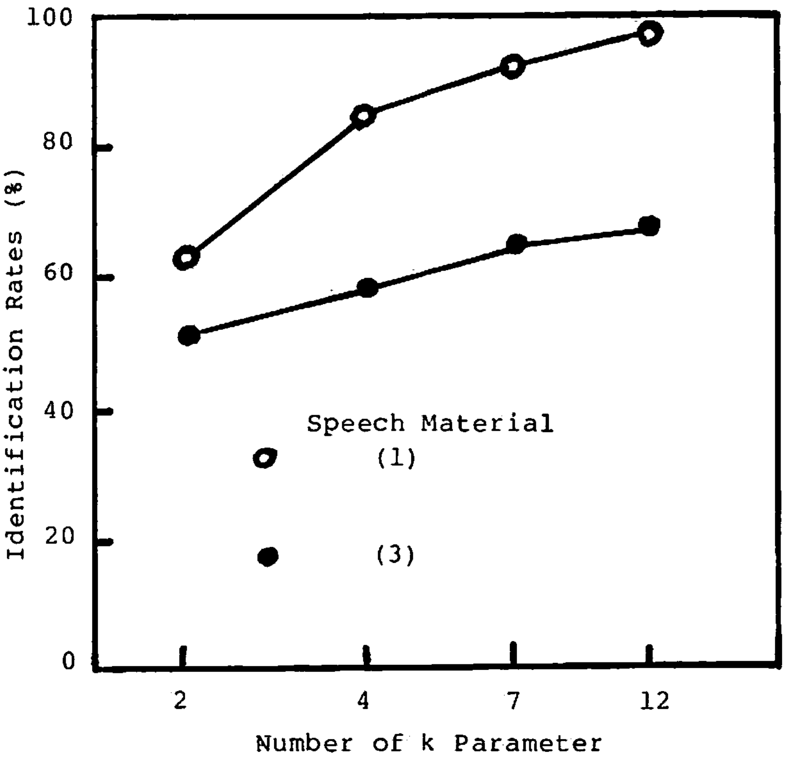

Fig. 3 Results of speaker identification test, Experiment 2

Experiment 3

The effects of the acoustic feature parameters on speaker identification were measured in Experiment 1 and it was concluded that the effect of the frequency spectrum envelope was prominent. To evaluate quantitatively the contribution of the frequency spectrum envelope and the excitation source signal on speaker identification, a third experiment was executed, in which the feature parameters of two speakers were combined to synthesize one test signal. Such a compound test signal was reproduced with two PARCOR analyzers and one synthesizer.

In this experiment, speech material (1) was not used as an input speech signal of the PARCOR analyzer. Even if the speech material (2) was used for synthesizing the compound test signal, some disagreement of temporal characteristics remain between the feature parameters of different speakers. Such disagreement is removed by the use of a dynamic programming matching procedure.

An identification test signal was presented to the listener using both the "naming" and the ABX methods. Results are shown in Fig. 4 (a) and (b). In these figures, the identification rate allotted to the speaker of the frequency spectrum envelope is denoted as S, that of the excitation source signal as E and that allotted to another speaker is A. In the "naming" method, identification rates allotted to another speaker were about 13 and

24 % for speech materials (2) and (3), respectively. In the ABX methods, such identification rates are replaced with those allotted to the speaker of the excitation source signal. It may be concluded that (1) the contribution of the frequency spectrum envelope is superior and more stable for speaker identification than the excitation source signal, (2) the contribution ratios of the frequency spectrum envelope to the excitation source signal are about 4:1 and 2:1 for speech materials (2) and (3), respectively.

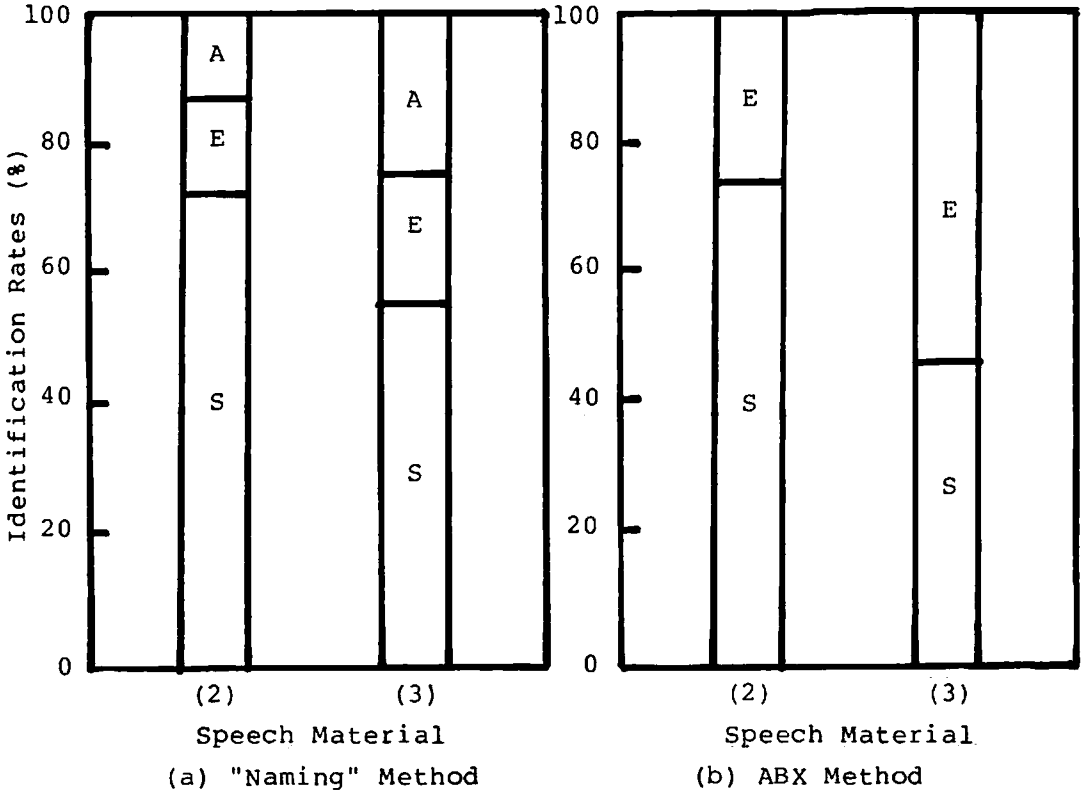

Fig. 4 Results of speaker identification test, Experiment 3.

Conclusion

Summarizing the three experiments on auditory identification of speaker, it may be concluded that:

(1) Among acoustic feature parameters, the frequency spectrum envelope has the greatest effect on speaker identification. The excitation source signal in addition to the temporal characteristics of the utterance have an effect on speaker identification, only if the frequency spectrum envelope of the speech is removed from the speech signal.

(2) Distortion of the frequency spectrum envelope as caused by the reduction of the number of k parameter significantly affects the dynamic characteristics of speech and causes deterioration of the speaker identification rate.

(3) The contribution ratio of the frequency spectrum envelope to the

excitation source signal in auditory identification of speaker is about 4:1 in conversational speech.

References

(1) I. Pollack, J.M. Pickett and W. H. Sumby, "On the identification of speakers by voice", J.A.S.A., 26, p. 403, 1954.

(2) A.J. Compton, "Effects of filtering and vocal duration upon the identification of speakers, aurally", J.A.S.A., 35, p. 1748, 1963.

(3) P.D. Bricker and S. Pruzansky, "Effects of stimulus content and duration on talker identification", J.A.S.A., 40, p. 1441, 1966.

(4) W.D. Voiers, "Perceptual bases of speaker identity", J. A.S.A., 36, p. 1065, 1964.

(5) K.N. Stevens, C.E. Williams, J.R. Carbonell and B. Woods, "Speaker authentication and identification: a comparison of spectrographic and auditory presentations of speech material ", J.A.S.A., 44, p. 1596, 1968.

(6) F.R. Clarke and R.W. Becker, "Comparison of techniques for discriminating among talkers", J.Speech and Hearing Res., 12, p. 747, 1969.