Ann. Bull. RILP, No. 14, 209-216 (1980)

EFFECT OF ACOUSTIC FEATURE SPEECH PARAMETERS ON PERCEPTUAL IDENTIFICATION OF SPEAKER

Shuzo Saito and Kenzo Itoh*

1. Introduction

Several studies have previously reported on the auditory identification of speaker with reference to the effect of the frequency band, duration time fundamental frequency of speech signal, etc. [1]-[3] Studies have also been done to estimate the effects of feature speech parameters on auditory identification of the speaker using monosyllable and vowel speech materials.[4], [5] In the present study, auditory speaker identification measurements have been performed to estimate the speaker-dependent feature of speech by the use of the synthesized speech sounds of sentences and sustained vowels, processed by the PARCOR speech analysis and synthesis method.

2. Test Procedure

The speech signal was passed through a low pass filter of 3.4kHz cutoff frequency, and its amplitude was then digitized into 12 bits at intervals of 125 microseconds. The digitized signal was then fed to the PARCOR analyzer, and the k parameters, speech power, ratio of voiced and unvoiced speech powers and the fundamental frequency of voice excitation were extracted for every frame period of five milliseconds. These analyzed parameters were then partially modified and fed to the PARCOR synthesizer to produce various kinds of test sounds used for auditory identification of speaker.

Three kinds of speech materials were used for the speaker identification test. The first was the utterance of a sentence having a duration time of about 3 seconds (Material 1). The second was the utterance of the same sentence, but the timing of word spurts in the sentence was forced to mimic the utterance of a specified speaker (Material 2). The third was the utterance of the five Japanese vowels /a, i, u, e, o/, where each was extracted in 335 milliseconds from the utterances of the sustained vowels; the five vowels were presented successively with silent intervals of 67 milliseconds (Material 3). The speech materials were uttered by five male speakers. Twelve male listeners with normal hearing were engaged in the speaker identification tests in a monaural listening condition. All listeners were colleagues of the five speakers and were familiar with their voices.

Various kinds of synthesized speech signals were presented in random order, and the listeners were requested to identify the speaker for each synthesized signal, namely, the 'naming' method was used in most identification

* Musashino Electrical Communication Laboratory, Nippon Telegraph and Telephone Public Corporation.

tests. In a few tests, the ABX method was used supplementally, where listeners were requested to judge whether the speaker of X signal was more likely to be the speaker of signal A or B.

3. Results of Speaker Identification Experiments

Experiment 1: Effect of the frequency spectrum envelope on speaker identification

The effect of the frequency spectrum envelope on speaker identification was tested for two kinds of speech materials, (1) and (3), changing the number of k parameters of the PARCOR synthesizer to 12, 7, 4, 2 and 0. To check the effect of the fundamental frequency of voice excitation, two conditions of fundamental frequencies were used for speech material (1), that is, one was the same as analyzed, and the other was fixed at 113Hz, which was the average value of all speakers. The 'naming' method was used for listener response. The result is shown in Fig. 1. In this figure, the speaker identification rates of the original speech signal, that is, the nonsynthesized signal, are also shown as AD/DA of the abscissa.

It is seen that the effects of the frequency spectrum envelope are rather similar for both speech materials (1) and (3), although there is an inherent reduction of the identification rate in speech material (3), caused by the absence of the temporal feature of speech. The identification rate of the fixed pitch signal is comparable to that of the natural pitch signal in the increased k parameter condition, but it becomes close to that of the sustained vowels in the decreased k parameter condition. It is assumed that the effect of the temporal variation of fundamental frequency on speaker identification becomes remarkable corresponding to the reduction of k parameter, that is, the deterioration of the frequency spectrum envelope.

Experiment 2: Effect of the excitation source signal on speaker identification

The effect of the excitation source signal on speaker identification was tested for speech materials (1) and (2), modifying the feature parameters of speech. Two kinds of modification technique were used: one fixing the fundamental frequency of vocal excitation at 113Hz (fixed pitch), and the other deleting the frequency spectrum envelope from the synthesized speech (excitation only). Next, four kinds of test signals were derived as follows: (a) unmodified speech, (b) fixed fundamental frequency, (c) excitation source signal only, and (d) excitation source signal accompanied by fixed pitch. These were tested by the 'naming' method for listener response.

The result is shown in Fig. 2. Four kinds of modification for the test signals are represented in the abscissa. It is seen that the difference between the identification rates of (d) and (e) is significant, while that of (b) and (c) is negligibly small. It may therefore be concluded that the effect of the frequency spectrum envelope is more essential and significant than that of the excitation source signal for speaker identification, and the latter becomes significant provided that the frequency spectrum envelope is deleted from the test speech signal.

Experiment 3: Effect of the timing of word spurts in a sentence on speaker identification

The effect of the timing of word spurts in a sentence was tested for speech material (2). The test procedure was the same as in Experiment 2, except for the choice of test speech material. The results are shown in Fig. 2. It is seen that the effect of the timing of word spurts is much like that of the excitation source signal: it becomes significant when the frequency spectrum envelope is deleted from the speech signal. It is noticed that the reduction of the identification rate caused by modifying the timing of word spurts is comparable to that of the fixed pitch as described in Experiment 2, and these two features together affect on speaker identification.

Experiment 4: Effect of the temporal variation of the feature parameter

To evaluate the effect of the temporal variations of the frequency spectrum envelope and the excitation source signal on speaker identification, the frame period of these feature parameters on the PARCOR synthesizer were varied from 5 to 80 milliseconds. Variation modes of the frame periods were not only individual, but also simultaneous in the two feature parameters. Both the unmodified and the fixed pitch test signals of speech material (2) were used for this experiment. The 'naming' method was used for listener response.

The results are shown in Figs. 3(a) and (b), for the unmodified and the fixed pitch signals, respectively. In these figures, the individual effects of the temporal variations of the feature parameters are expressed by the symbols S and E for the frequency spectrum envelope and the excitation source signal, respectively, and the effect of the simultaneous variations of the two feature parameters is denoted as S + E. It is seen that the effect of the frame period length is valid in the simultaneous variations of the feature parameters as shown in Fig. 3(a). In the case of the fixed pitch test signal, the contribution of the excitation source signal is reduced, as is seen in Fig. 3(b). It is concluded that the effects of the temporal variations of the feature parameters complement each other in terms of speaker identification.

Experiment 5: Rates of contribution of the two feature parameters on speaker identification

To evaluate the contribution of the feature parameters on speaker identification quantitatively, a compound synthesis technique was used in which the feature parameters of two speakers were combined to synthesize one test signal. This technique is rather similar to that used by J. E. Miller.[4] In this experiment, speech materials (2) and (3) were used to synthesize the compound test signal. Several slips remained between the feature parameters of different speakers in speech material (2). Such slips were removed by a dynamic programming matching procedure. The test signal was presented in both the 'naming' and ABX methods.

The result is shown in Fig. 4. In this figure, 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

as A. It is seen that the errors allotted to another speaker in the 'naming' method are mostly transferred to the speaker of the excitation source signal in the ABX method. It may be concluded that the contribution of the frequency spectrum envelope is more effective than the excitation source signal, and the ratios of contribution of the former to the latter on speaker identification are about 4:1 and 2:1 for speech materials (2) and (3), respectively.

4. Discussion

The effects of feature parameters of speech signals on auditory identification of a speaker have been estimated using sentence and vowel speech materials. It has been found that there are significant differences in the identification rates between a sentence and sustained vowels, which stems from inherent dynamic characteristics. To check the effect of the dynamic characteristics on speaker identification, a supplementary test was performed. A monosyllable /ba/ was used for the speaker identification test, in addition to a sentence and sustained vowels. The results are shown in Fig. 5. In this figure, the identification rates of three speech materials are shown as white bars, and the rates of the monosyllable and the vowels reproduced in the inverse direction are in hatched bars. It is seen that the identification rate of the monosyllable lies at the middle of the other speech materials, and it is reduced significantly in the case of the inverse reproduction. It may be concluded that the dynamic characteristics of speech signal play an important role even in the case of the monosyllable.

5. Conclusion

Summarizing our experiments on auditory identification of a speaker, it is concluded as follows:

(1) The effect of the frequency spectrum is prominent among the feature parameters of speech on auditory identification of speaker.
(2) The effects of the fundamental frequency and the timing of word spurts become significant, provided that the frequency spectrum envelope is removed from the original speech signal.
(3) The effects of the temporal variations of the feature parameters of speech are complementary to each other on auditory identification of the speaker.

References

(1) Pollack, I., J.M. Pickett and W.H. Samby (1954); On the Identification of Speaker by Voice. J. Acoust. Soc. Amer., 26, p.403.
(2) Bricker, P.D. and S. Pruzansky (1966); Effect of Stimulus Content and Duration on Talker Identification. J. Acoust. Soc. Amer., 40, p. 1441.
(3) Compton A.J. (1963); Effect of Filtering and Vocal Duration upon the Identification of Speakers, Aurally. J. Acoust. Soc. Amer., 35, p. 1748.

(4) Miller, J. E. (1964); Decapitation and Recapitation, a Study of Voice Quality. J. Acoust. Soc. Amer., 36, p. 2002 (A).
(5) Wood, C. A. (1977); Source/Vocal Tract Influence on Speaker Discrimination, 9th ICA, Madrid. p. 497.

Acknowledgment

This study was supported in part by a Grant in Aid for Scientific Research No. 540003 from the Japanese Ministry of Education, Science and Culture.

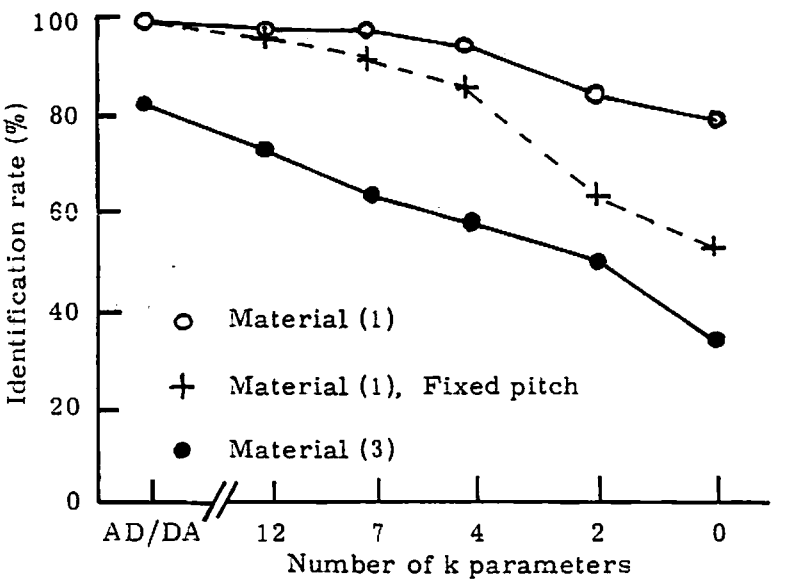

Fig. 1 Effect of the number of k parameters on speaker identification

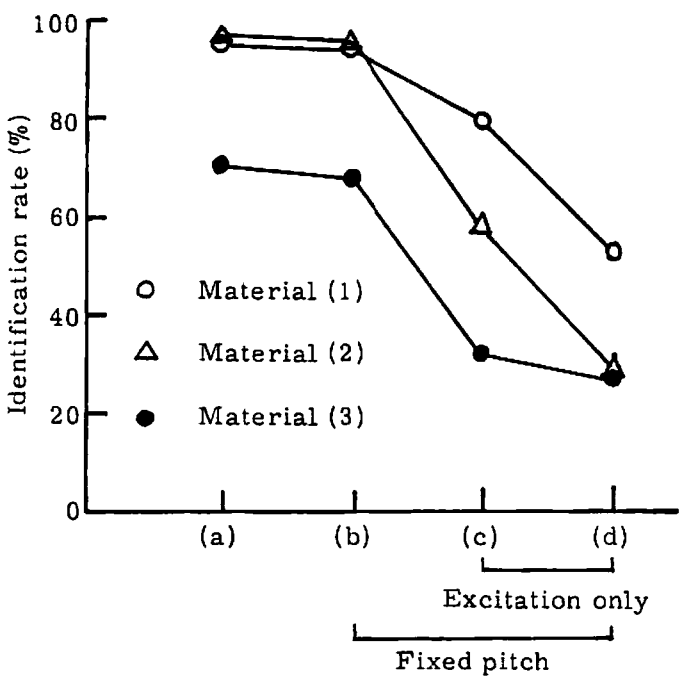

Fig. 2 Effect of the excitation source signal on speaker identification

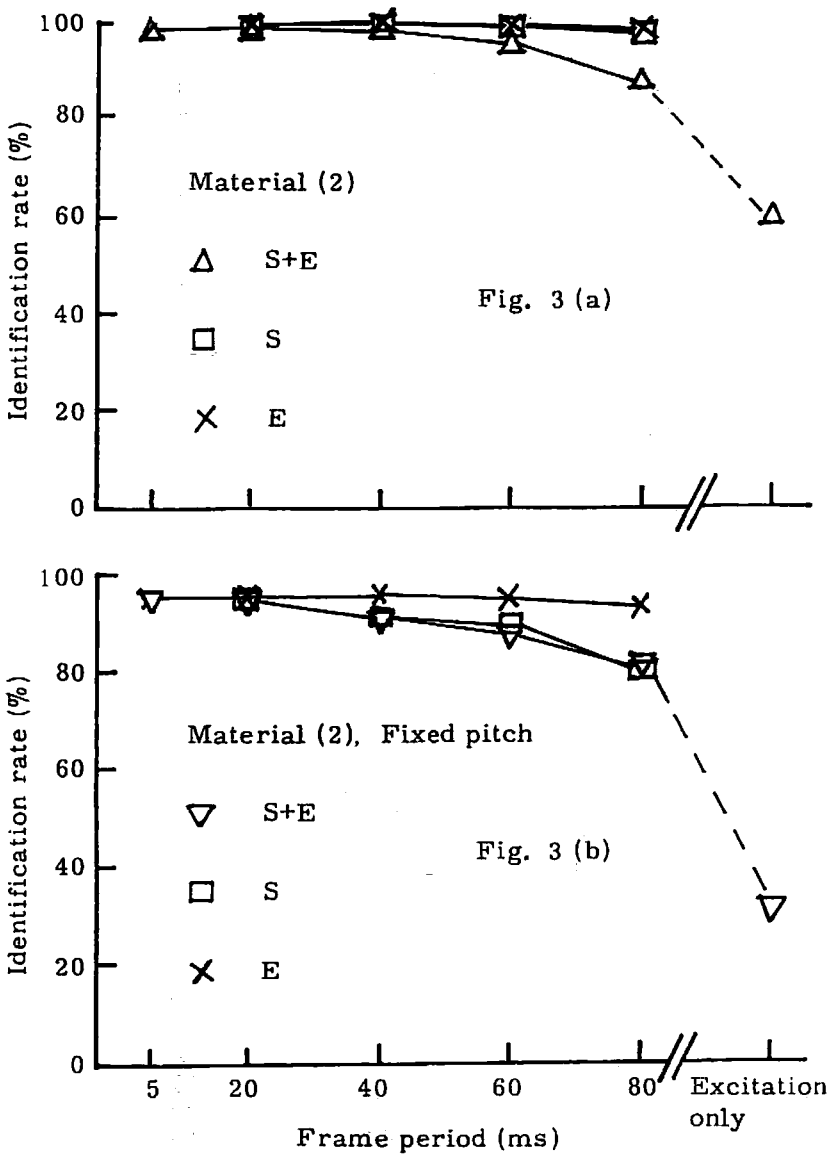

Fig. 3 Effect of the frame period of the feature parameter on speaker identification

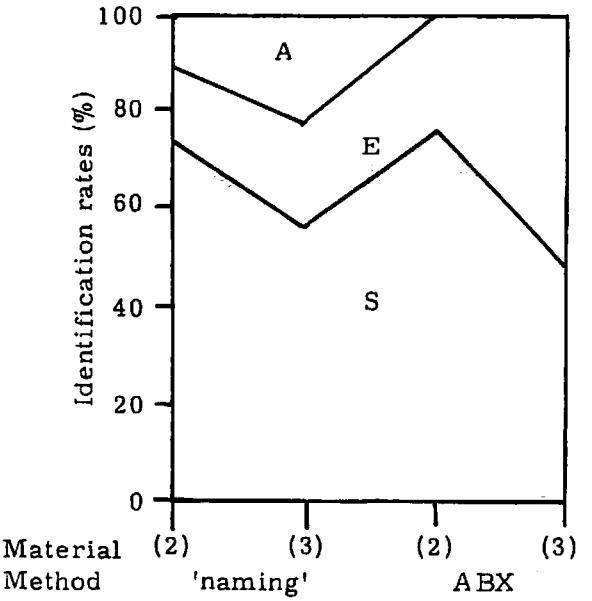

Fig. 4 Contribution rates of the feature parameters on speaker identification

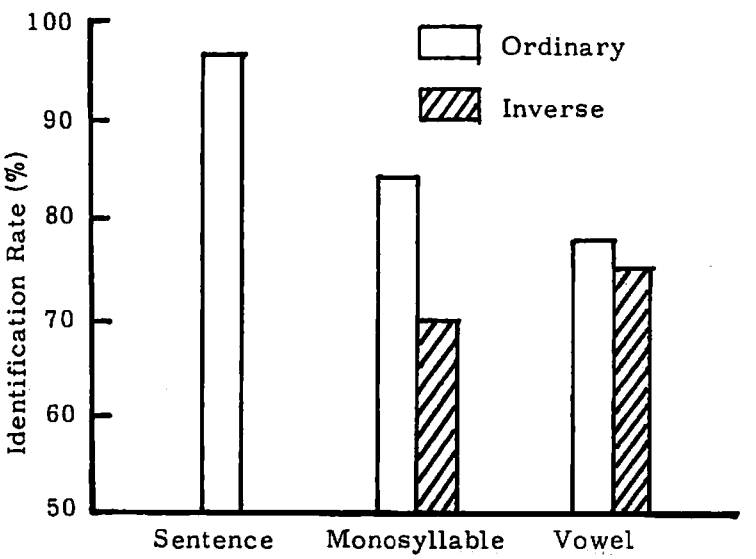

Fig. 5 Effects of the speech materials and the inverse reproduction on speaker identification