Analysis of acoustic features affecting speaker identification

Norio Higuchi and Makoto Hashimoto

ATR Interpreting Telecommunications Research Laboratories,
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-02 Japan

(Received 3 July 1995)

Keywords: Speaker identification, Predictive model, Perceptual contribution, Spectral information, Prosodic information

PACS number: 43.71.Bp

1. Introduction

Realizing variety in synthetic speech quality is one of the most important research issues in speech synthesis as well as the improvement of naturalness. A few previous studies investigated variation due to speaking style and speaker individuality. Achieving the variation of speaker individuality is useful not only for satisfying the listener's preference but also for speaker identification in a speech translation system for teleconferencing. For speaker individuality in Japanese, a previous study showed that spectral information was dominant over prosodic information. However, in this study subjects were acquainted with the speakers. This is not a common condition for speaker identification, and there was no quantitative consideration given to the amount of difference among the acoustic features.

In order to quantitatively investigate the influence of individual acoustic features, we have proposed a new model to predict the contribution rate of each acoustic feature in the human process of speaker identification based on the result of a hearing test using speech resynthesized by swapping the acoustic features. The model can predict the contribution rates of spectral, mean logarithmic fundamental frequency and other features with an error of 7.71%.

2. Assumption for the modeling

The following three assumptions were used in making a model for the prediction of the contribution rate of each acoustic feature.

(1) By swapping an acoustic feature of Speaker B with one of Speaker A, the rate of judgment that resynthesized speech is closer to Speaker A than to Speaker B increases and the amount of this increase does not depend on the condition of other acoustic features.

(2) The total acoustic difference between samples spoken by two different speakers is equal to the summation of the acoustic differences in all of the acoustic features.

(3) The ratio of acoustic difference of each acoustic feature is proportional to that of the contribution rate in a perception test of speaker identification.

The contribution rate of each acoustic feature was measured in a perception test of speaker identification, and this test's result was used to optimize the model to predict the contribution rate based on the difference in each acoustic feature. The predictive error of the contribution rate was evaluated. Figure 1 diagrams these procedures.

3. Speech samples

The three short sentences spoken by six male speakers in the ATR Database of Continuous Speech were used here. One speaker was assigned as the standard

Fig. 1 Flow of the experiment.

Fig. 2 Speech synthesized by swapping acoustic features.
speaker, and three pairs of sentence speech samples spoken by the standard speaker and five other speakers were used for a hearing test by swapping each acoustic feature. Total number of sample pairs is fifteen.

4. Measurement of perceptual contribution rate

In order to measure the perceptual contribution rate of each acoustic feature, acoustic features were divided into three categories: fundamental frequency, spectral information and phoneme durations. Eight kinds of speech were resynthesized by swapping among these three categories (Fig. 2) and used in an A-B-X test, in which subjects judge whether the synthetic speech is closer to Speaker A or Speaker B. The number of subjects was nine, and each synthetic sample was judged four times by each subject. The contribution rate of each acoustic feature was derived from the following three equations based on the result of a hearing test.

Contribution rate of fundamental frequency
\[ \frac{1}{4} \sum [P_A(A, X, Y) - P_B(B, Y, Z)] \]

Contribution rate of spectral information
\[ \frac{1}{4} \sum [P_A(A, X, Y) - P_A(X, B, Z)] \]

Contribution rate of phoneme durations
\[ \frac{1}{4} \sum [P_A(A, X, Y, Z) - P_A(X, Y, B)] \]

where \( P_A(X, Y, Z) \) is the probability that the synthetic speech whose mean fundamental frequency, spectrum and phoneme durations are equal to those of Speaker X, Speaker Y and Speaker Z, respectively, was judged to be closer to Speaker A than to Speaker B. X, Y and Z may each be A or B.

5. Predictive model for perceptual contribution rate

Figure 3 represents the relationship between acoustic difference and the perceptual contribution rate of each acoustic feature. Total acoustic difference is the summation of the differences in all of the acoustic features (Assumption 2). The ratio of perceptual contribution rates is proportional to that of the amount of difference in each acoustic feature (Assumption 3). The function for the prediction of the perceptual contribution rates is too complicated to optimize weighting factors by solving equations; therefore, optimization was performed by an Analysis-by-Synthesis Method. Measures of the difference in each acoustic feature are mean logarithmic fundamental frequency, cepstral distance and mean mora duration.

6. Evaluation of predictive error

The result of a preliminary test makes it obvious that mean mora duration is not sufficient for describing the temporal difference, so a constant value 1 was used to represent the difference in other acoustic features, which includes the difference in phoneme duration of each phoneme and power. Figure 4 shows the result after optimizing the weighting factor of each acoustic feature. The best values of the weighting factors for the

![Fig. 3 Predictive model for contribution rate of each acoustic feature. (In the case of two acoustic features)](image-url)

![Fig. 4 Relationship between perceptual contribution rates obtained in a hearing test and their predicted values based on the model.](image-url)
7. Conclusions

In order to investigate the influence due to the acoustic features in the human process of speaker identification, the authors have proposed a new model for the prediction of the perceptual contribution rate of each acoustic feature; consequently, the predictive error is 7.71% in RMS after optimization of the weighting factor of each acoustic feature.

As the weighting factors are good indices of the mechanism of human speaker identification, this predictive model is effective in analyzing the mechanism of speaker identification in different cases, for example, the case where subjects are acquainted with speakers and other case where they aren't. It is also useful to evaluate distance measure of each acoustic feature itself based on the degree of fitting.

References


Norio Higuchi was born in Yamanashi, Japan, in 1954. He received the B. E., M. E. and Ph. D degrees from University of Tokyo, Tokyo, in 1977, 1979 and 1982, respectively. He joined Kokusai Denshin Denwa Co., Ltd. in 1982 and moved to ATR in 1993. He is the head of Department 2 in ATR Interpreting Telecommunications Research Laboratories, where he is directing research in speech synthesis, evaluation of synthetic speech quality, voice conversion and prosody interpretation. He received Awaya Kiyoshi award of Acoustical Society of Japan in 1985. Dr. Higuchi is a member of IEICE of Japan and IEEE.

Makoto Hashimoto was born in Kobe, Japan, in 1965. He received the B. E. degree from Ritsumeikan University, Kyoto, Japan, in 1989. He joined SANYO Electric Co., Ltd. in 1989, Osaka, Japan. Since 1993, he has been a researcher at ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan, where he has engaged in research in speech synthesis, especially voice conversion. His research interests include speech synthesis and speech recognition. He is a member of the Acoustical Society of Japan.