EFFECT OF SPEAKER FAMILIARITY AND BACKGROUND NOISE ON ACOUSTIC FEATURES USED IN SPEAKER IDENTIFICATION

Satoshi Kitagawa, Makoto Hashimoto† and Norio Higuchi

e-mail: satoshi@itl.atr.co.jp
ATR Interpreting Telecommunications Research Labs.
2-2 Hikaridai, Seika-cho, Soraku-gun, 619-02 Kyoto, Japan

ABSTRACT

In order to investigate the relationship between human perception in speaker identification and acoustic features (fundamental frequency ($f0$), spectrum, and duration) under various communication conditions, this paper describes several perception experiments and an approach to predict the perceptual contribution rate of each feature. Factors taken into account in this paper are: (1) speaker familiarity and (2) background noise. As a result, it is shown that: (1) the perceptual contribution rate increases as the distance of an acoustic feature increases, (2) the spectral contribution rates for familiar speakers are larger than those for unfamiliar speakers, (3) the contribution of $f0$ tends to increase as the noise increases, and (4) in case of the same S/N ratio, the contribution of $f0$ in the computer room noise environment is larger than in the car noise environment.

1. INTRODUCTION

Humans can differentiate speaker characteristics simply by hearing a person speak. We assume that several acoustic features ($f0$, spectrum, and duration) convey these individual speaker characteristics; therefore, it is very useful to investigate the contribution of each feature to the human perception process of speaker identification. In previous studies[?][?], we used the ATR database to show that the perceptual contribution rate of each acoustic feature depends on the difference of each acoustic feature between two speakers. A model was proposed that can predict the contribution rate with little error.

In this paper, we measured the contribution rates of acoustic features for both familiar speakers and unfamiliar speakers, and both noiseless speech and speech in noisy environments by hearing test. Then we analyzed the effect due to the difference among them based on the weighting factor of each acoustic feature which reflects the difference in a communication condition.

†Presently; SANYO Electric Co., Ltd.

2. EXPERIMENTAL CONDITIONS

In the experiment to analyze the effect of speaker familiarity, we used speech samples uttered by seven male speakers who work in the same department as the familiar speakers and those uttered by six male professional announcers and narrators as unfamiliar speakers which are included in ATR speech database[?]. There were nine subjects.

The speech samples uttered by six male announcers and narrators were used also in the experiment for the effect of background noise. Two kinds of noise were added for the noisy environment: "noise in the running car (the 2000cc class)" and "noise in the computer room (the work station)" from the JEIDA database[?]. There were eight subjects.

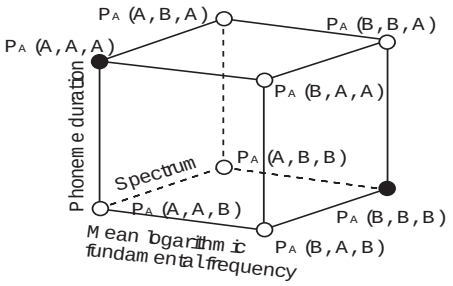

Figure 1: Speech samples resynthesized by swapping acoustic features.

In both experiments, six kinds of utterances were resynthesized by swapping three acoustic features shown in Fig.1, and the speech was used in an A-B-X test where the subjects judged whether the synthetic speech (X) was closer to Speaker A or Speaker B.

3. CONTRIBUTION RATE AND PREDICTION MODEL

Based on the results of a hearing experiment, the following calculation for the contribution rate of $f0$ (C_{f0}) was formulated;

$$C_{f0} = \frac{1}{4} \sum \{P_A(A,Y,Z) - P_A(B,Y,Z)\}$$

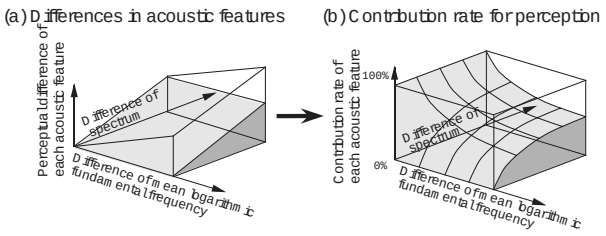

Figure 2: Predictive model for the contribution rate of each acoustic feature (In the case of two acoustic features).

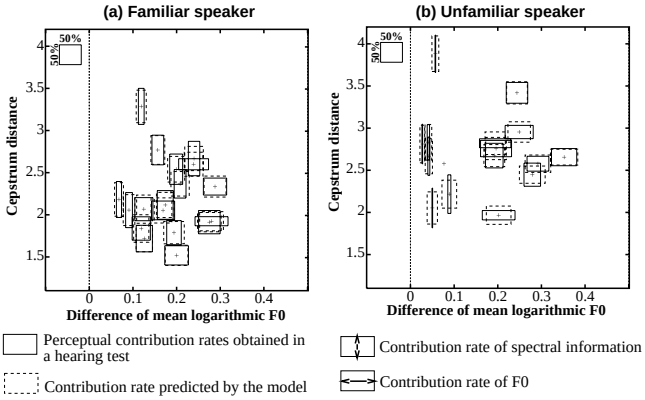

Figure 3: Relationship between the acoustic feature distance and the perceptual contribution rate for familiar and unfamiliar speakers.

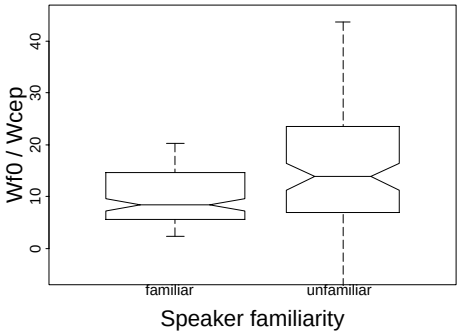

Figure 4: Relative weight of $f0$ to spectrum for familiar and unfamiliar speakers.

where $P_A(X, Y, Z)$ is the probability that the synthetic speech (with mean $f0$, spectrum and phoneme duration equal to those of Speaker X, Speaker Y and Speaker Z), is judged to be closer to Speaker A than to Speaker B. X, Y and Z may be either A or B. The perceptual contribution rates of spectrum (C_{spec}) and duration (C_{dur}) were also defined in the same way.

Figure 2 represents the relationship between the acoustic difference and the perceptual contribution rate of each acoustic feature. The total acoustic difference is a summation of the differences in all of the acoustic features. The ratio of perceptual contribution rates is proportional to that of the amount of difference in each acoustic feature. The function for predicting of the perceptual contribution rates is too complicated for optimizing the weighting factors by solving equations. Therefore, optimization was performed by an Analysis-by-Synthesis Method.

The contribution rate of $f0$ are defined as follows:

$$\hat{C}_{f0} = w_f \cdot D_{f0}/(w_f \cdot D_{f0} + w_s \cdot D_{spec} + w_d \cdot D_{dur})$$

where $\hat{C}_{f0}$ is the predictive contribution rate for $f0$, and D_{f0}, D_{spec} and D_{dur} are the distance of the acoustic feature for $f0$, spectrum and duration, respectively, and w_f, w_s and w_d are the weighting factor for the distance of each acoustic feature. The predictive contribution rates of spectrum ($\hat{C}_{spec}$) and duration ($\hat{C}_{dur}$) were also defined in the same way. The difference in $f0$ and spectrum was measured by using the mean logarithmic $f0$ and cepstral distance.

4. RESULTS

Experimental results show that the contribution rates of $f0$ and cepstrum were high and that the contribution rate of duration was low for each condition.

4.1. Speaker familiarity

Figure 3 shows the relationship between the acoustic feature distance and the contribution rate for $f0$ and cepstrum of familiar speaker and unfamiliar speaker. In Fig.3, the solid line shows the contribution rate obtained in the hearing test, and the broken line shows the rate predicted by the model.

Figure 3 indicates that the perceptual contribution rate increased as the acoustic feature distance increased; the contribution of cepstrum for the familiar speaker was larger than that for the unfamiliar speaker. By optimizing the weighting factors [w_f, w_s, w_d] in the prediction model, the prediction errors were minimum when:

[1.000, 0.122, 0.027] (familiar speakers)

[1.000, 0.079, 0.056] (unfamiliar speakers)

The prediction errors were 10.6% (familiar) and 13.4% (unfamiliar) in RMS. Based on the optimized weighting factor, the contribution of spectrum for the familiar speakers was about 1.5 (=0.122/0.079) times that for the unfamiliar speakers.

Next an analysis by a statistical technique was carried out to examine the significance of the difference between familiar and unfamiliar speakers. Figure 4 shows the distribution of the ratio of the weighting factor for $f0$ to that for cepstrum. In this figure,

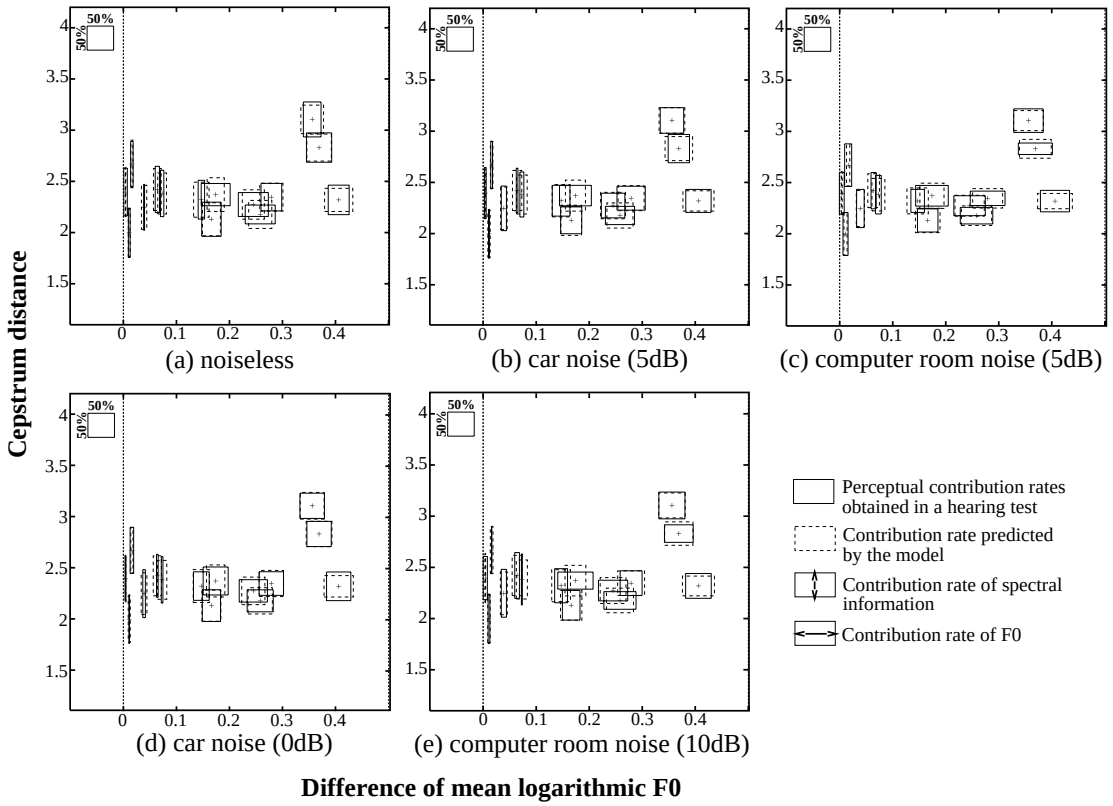

Figure 5: Relationship between the acoustic feature distance and the perceptual contribution rate for noiseless and noisy condition.

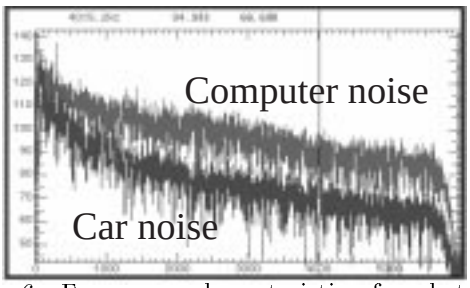

Figure 6: Frequency characteristic of each type of noise

the notch shows the 95% confidence interval; and a bigger ratio indicates a greater distribution of the fundamental frequency. In Fig.4, it is significant that the contribution of spectrum for speaker identification to familiar speakers is greater than that to unfamiliar speakers.

4.2. Background noise

Figures 5(a)–(e) show the relationship between the acoustic feature distance and the contribution rate for $f0$ and cepstrum; for an S/N ratio = ∞, 5dB (car), 5dB (computer room), 0dB (car) and 10dB (computer room), respectively.

Figure 5 indicates that the contribution rate of each acoustic feature depends on the amount of the acoustic difference between a speaker pair. By optimizing the weighting factors $[w_f, w_s, w_d]$ in the prediction model, the prediction errors were minimum when:

$[1.000, 0.149, 0.077]$ (S/N ratio = ∞ dB)
$[1.000, 0.118, 0.074]$ (5dB, car)
$[1.000, 0.080, 0.177]$ (5dB, computer room)
$[1.000, 0.133, 0.136]$ (0dB, car) and
$[1.000, 0.112, 0.050]$ (10dB, computer room).

And the prediction error in each case is 9.0%, 7.5%, 8.2%, 7.0% and 9.9% in RMS, respectively. Based on the optimized weighting factor, the contribution of spectrum decreases by the ratios shown below when any kinds of noise were added:

0.78 (=0.118/0.149), [5dB, car]
0.54 (=0.080/0.149), [5dB, computer room]
0.89 (=0.133/0.149), [0dB, car] and
0.75 (=0.112/0.149), [10dB, computer room].

In order to analyze the change of contribution rates due to the additional noise, we plotted the contribution rates of $f0$ and spectrum for both noiseless

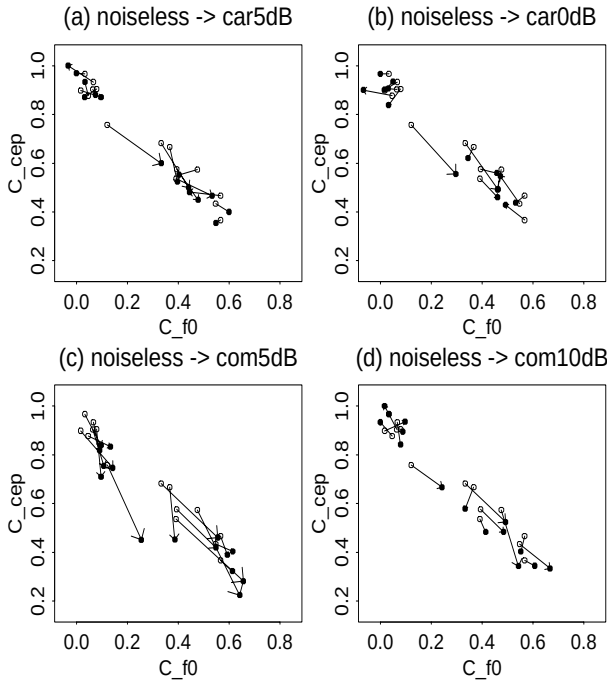

Figure 7: Change in the contribution rate due to background noise

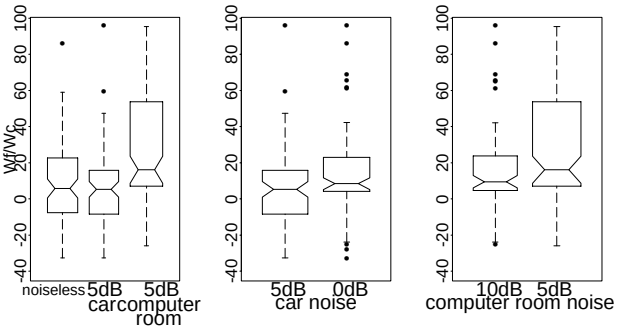

Figure 8: Relative weight of $f0$ to spectrum for noiseless and noisy conditions.

and noisy conditions. The arrow indicates the change due to additional noise (see Fig.7). It shows that the contribution of $f0$ for speaker identification in the noisy environment was greater than in the noiseless environment.

The relative weight of $f0$ to spectrum for noiseless and noisy conditions was shown in Fig.8. The figure shows that

- The contribution of $f0$ tends to increase as the noise increases.
- In the case of the same S/N ratio, the contribution of $f0$ in the computer room noise environment is larger than in the car noise environment.

This is probably because computer room noise masks the spectral information of speech in a boarder band than car noise, such that the contribution of $f0$ under the computer room noise becomes larger than under the car noise.

5. CONCLUSIONS

We analyzed the importance of acoustic features affecting speaker identification in various communication conditions.

Several perception experiments were carried out to measure the contribution rate in speaker identification of the fundamental frequency, spectrum and duration.

In addition, a prediction model for the perceptual contribution rate was constructed and evaluated in term of prediction errors in the results of a hearing experiment.

The following results were obtained;

- The perceptual contribution rate increases as the distance of the acoustic feature between a speaker pair increases.
- The contribution rate of spectrum for familiar speakers is larger than that for unfamiliar speakers.
- In the case of the same S/N ratio, the contribution of $f0$ in the computer room noise environment is larger than in the car noise environment.
- The contribution of $f0$ tends to increase as the noise increases.
- The prediction errors of the prediction model for the contribution rates were 7.0–13.4%; it means that the model can estimate the contribution rate with small prediction error.

In this paper, we examined speaker identification for Japanese read speech. In the future, we want to analyze contribution rates for the utterances of other languages and other speaking styles.

REFERENCES

[1] N. Higuchi and M. Hashimoto : Proc. of EUROSPEECH'95, pp.435–438 (1995.9).

[2] N. Higuchi and M. Hashimoto : J. Acoust. Soc. Jpn(E), **17**, pp.33–35 (1996.1).

[3] M.Abe, Y.Sagisaka, T.Umeda and H.Kuwabara : "Speech Database User's Manual", Tech. report of ATR, TR-I-0166 (1990)(in Japanese).

[4] Speech Input/Output Systems Expert Commitee Japan Electronic Industry Development Association: "JEIDA NOISE DATABASE", (1996).