<table>
<thead>
<tr>
<th>Title</th>
<th>Speaker individuality in fundamental frequency contours and its control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Akagi, Masato; Ienaga, Taro</td>
</tr>
<tr>
<td>Citation</td>
<td>Journal of the Acoustical Society of Japan, 18(2): 73-80</td>
</tr>
<tr>
<td>Issue Date</td>
<td>1997</td>
</tr>
<tr>
<td>Type</td>
<td>Journal Article</td>
</tr>
<tr>
<td>Text version</td>
<td>publisher</td>
</tr>
<tr>
<td>URL</td>
<td><a href="http://hdl.handle.net/10119/4884">http://hdl.handle.net/10119/4884</a></td>
</tr>
</tbody>
</table>
Speaker individuality in fundamental frequency contours and its control

Masato Akagi and Taro Ienaga

School of Information Science, Japan Advanced Institute of Science and Technology, 1-1, Asahidai, Tatsunokuchi, Ishikawa, 923-12 Japan

(Received 27 July 1996)

Speaker individualities in fundamental frequency ($F_0$) contours are investigated through analyses of several speakers' uttered speech and psychoacoustic experiments. The analyses are performed to extract significant physical characteristics of $F_0$ by using Fujisaki and Hirose's analysis method and the $F$-ratio of each physical characteristic. The experiments are performed to clarify the relationship between these physical characteristics and the perception of speaker's speech. The stimuli used in the experiments are re-synthesized with manipulated $F_0$ contours and spectral envelopes averaged overall for all speakers by using the Log Magnitude Approximation analysis-synthesis system. The analysis and experimental results indicate that (1) there is speaker individuality in the $F_0$ contours, (2) some specific parameters related to the dynamics of $F_0$ contours have many speaker individuality features and speaker individuality can be controlled by manipulating these parameters, and (3) although there are speaker individuality features in the time-averaged $F_0$, they help improve speaker identification less than the dynamics of the $F_0$ contours.

Keywords: Features of speaker individuality, Speaker individuality control, Fundamental frequency contour

PACS number: 43. 70. Gr, 43. 71. Bp, 43. 72. Ja

1. INTRODUCTION

Speech is one of the most natural and useful means of communication for human. If it can be used between humans and machines, interaction between them may be able to be improved. Text-to-speech synthesis techniques play an important role in the communication process between humans and machines. So far, research has focused on improving articulation in speech synthesis. However, adding and controlling physical correlates of speaker individuality has become a significant problem for improving speech quality in speech synthesis. Adding speaker individualities to synthesized speech makes the speech sound more natural and easier to listen to. Physical characteristics related to speaker individuality must be specified before we can add and manipulate speaker individuality.

It is well known that speech has speaker individuality. However, it is not clear what physical characteristics of speech are related to speaker individuality. Although speech can be described by its physical aspects; that is fine spectral envelopes reflecting vocal tract features and pitch frequencies related to glottal vibration characteristics, the physical correlates of speaker individuality embedded in these physical aspects have not been discussed in detail.

Previous research has suggested that time-averaged spectral envelopes of the 2.5- to 3.5-kHz frequency band and time-averaged fundamental frequency ($F_0$) are mainly related to speaker individuality. The relationship between the peaks of spectral envelopes and speaker individuality has also been reported. These studies, however, have only addressed static characteristics such
as time-averaged spectral envelopes and \( F_0 \)\(^3\) or global shifts of formants and formant band-widths,\(^4\) but not dynamic property. Physical characteristics in actual speech deviate usually and the dynamics of these characteristics could become significant for controlling speaker individuality.

In this paper, we assume that the physical characteristics used by humans to identify speakers are significant physical characteristics representing speaker individuality. Since \( F_0 \) contours are significant factors in Japanese and are varied for speaker’s properties such as vocal organs, home districts and expressions of emotions, regarding these speaker’s properties as speaker individuality, we investigate the physical characteristics embedded in the \( F_0 \) contours related to speaker individuality. The physical characteristics embedded in the \( F_0 \) contours of words can be described through psychoacoustic experiments.

We used the analysis method proposed by Fujisaki and Hirose\(^5\) (Fujisaki \( F_0 \) model) to extract and manipulate the physical characteristics of \( F_0 \), and the Log Magnitude Approximation (LMA) analysis-synthesis system\(^6\) to prepare synthesized stimuli with varied \( F_0 \) contours.

Three psychoacoustic experiments, Experiment 1, 2 and 3, were performed in that order to clarify the following three questions; (1) whether speaker individuality still exists in the \( F_0 \) contours when spectral envelopes and amplitude contours are averaged and when \( F_0 \) contours are modeled by the Fujisaki \( F_0 \) model (Experiment 1), (2) whether some parameters of the Fujisaki \( F_0 \) model are effective in identifying speakers and whether perceived results of subjects are changed if these parameters are exchanged with those of other speakers (Experiment 2), and (3) whether speaker individuality can be manipulated by shifting the time-averaged \( F_0 \) frequency (Experiment 3).

Experimental results suggest that (1) speaker individualities exist in the \( F_0 \) contours; (2) some specific parameters representing \( F_0 \) contours have speaker individuality features, and the manipulation of these parameters can control speaker individualities; and (3) shifting the time-averaged \( F_0 \) frequency often used for automatic speaker identification or verification has little effect on speaker identification rates in psychoacoustic experiments.

2. ANALYSIS OF \( F_0 \) CONTOURS

To discuss the relationship between \( F_0 \) contours and the speaker individuality embedded in them, we need a method for representing \( F_0 \) contours. For this study we adopt the Fujisaki \( F_0 \) model,\(^5\) because it describes \( F_0 \) contours as consisting of two elements, phrase and accent, which can be controlled independently. Moreover, the number of parameters for describing \( F_0 \) contours is not many, and this method is generally used in Japanese text-to-speech applications.

In this chapter, \( F_0 \) contours of 3 mora words are analyzed by the Fujisaki \( F_0 \) model and the parameters obtained are compared to estimate which ones are related to speaker individuality.

2.1 Representation of \( F_0 \) Contours

The Fujisaki \( F_0 \) model represents \( F_0 \) contour \( F_0(t) \) as follows:\(^7\):

\[
\ln F_0(t) = \ln F_0 + \sum_{i} A_{pi} G_p(t - T_{pi}) \\
+ \sum_{j} A_{aj} (G_a(t - T_{aj}) - G_a(t - T_{aj})) \\
+ A_{pe} G_p(t - T_{pe}) \tag{1}
\]

\[
\begin{align*}
G_p(t) &= at \exp(-at) \\
G_a(t) &= \min[1 - (1 + \beta t) \exp(-\beta t), 0.9], \quad t \geq 0
\end{align*}
\]

where

- \( F_0 \) is the baseline value of an \( F_0 \) contour,
- \( A_{pi} \) is the magnitude of the \( i \)-th phrase command,
- \( A_{aj} \) is the amplitude of the \( j \)-th accent command,
- \( I \) is the number of phrase commands,
- \( J \) is the number of accent commands,
- \( T_{pi} \) is the timing of the \( i \)-th phrase command,
- \( T_{aj} \) and \( T_{aj} \) are the onset and offset of the \( j \)-th accent command, and
- \( \alpha \) and \( \beta \) are natural angular frequencies of the phrase and accent control mechanism, respectively.

Parameters \( \alpha \) and \( \beta \) characterize dynamic properties of the laryngeal mechanisms for phrase and accent control, and thus may not vary widely between utterances and speakers.\(^7\) They were fixed at \( \alpha = 3.0 \) and \( \beta = 20.0 \). The negative phrase command at the end of the utterance was used as \( T_3 \). A schematic figure of the Fujisaki \( F_0 \) model is shown in Fig. 1.

2.2 Speech Data

The speech data used for all of the experiments
were three-mora words with accented second mora: "aōi" (blue), "nagāi" (long), and "nīōu" (smell). Each word was uttered ten times by three male speakers: KI, KO, and YO. Although these speakers come from the Tohoku area (north of Japan), Tokyo area, and Hokuriku area (western Japan) areas, respectively, they speak standard Japanese usually. When recording the speech data, the speakers were instructed to utter the words with standard Japanese accent and without emotions.

Aspects of \( F_0 \) contours of words uttered by these speakers were different a little each other, especially \( F_0 \) contours of KI are relatively flatter than those of others. Although this feature may be related to speaker's home districts, regarding difference of speaker’s home districts as one of speaker individualities in this paper, we use these types of \( F_0 \) contours as the speech data.

The speech data were sampled at 20 kHz with 16-bit accuracy, and analyzed using 16-th order LPC in 30 ms Hanning window at every 5 ms period. The auto-correlation of their residuals was calculated to estimate their \( F_0 \) contours. Equation (1) was then fitted to the contours by the analysis-by-synthesis method. Figure 2 illustrates an \( F_0 \) contour and a fitted curve of the word “aōi” (blue) uttered by one male speaker.

2.3 Analysis Results

To identify physical characteristics that represent speaker individuality in the analyzed parameters, we calculated the \( F \) ratio (inter-speaker variation divided by averaged intra-speaker variations) for each parameter:

\[
F_k = \frac{\sum_i (c_{ik} - \frac{1}{N} \sum_i c_{ik})^2}{\left[ \frac{1}{M} \sum_j \sum_i (c_{ijk} - \bar{c}_{ik})^2 \right]^{\frac{1}{2}}},
\]

where \( c_{ijk} \) is the \( j \)-th observation of the \( i \)-th speaker for the parameter \( k \), \( M \) is the number of the observations and \( N \) is the number of the speakers. The largest \( F \) ratio indicates the parameter whose inter-speaker variation is large and whose intra-speaker variation is small, and suggests the most significant parameter for speaker identification.

Figure 3 shows the \( F \) ratio for each parameter of
the word "aōi" as an example. In the figure, $\Delta T_i$ indicates the difference between the command timing and the corresponding mora boundary. The aspects of the $F$ ratio of the other two words are almost the same.

These results indicate that the $F$ ratios of three parameters, $F_0$, $A_p$, and $A_d$, are much larger than those of the other parameters and suggests that the three parameters are significant for perceptual speaker identification. Parameter $F_0$ is related to a time-averaged $F_0$ frequency and parameters $A_p$ and $A_d$ are related to a dynamic range of an $F_0$ contour.

The time-averaged $F_0$ frequency is usually used in automatic speaker identification and verification systems. Thus, it is reasonable that the parameter $F_0$ should be significant for perceiving a speaker's speech. However, the dynamics of $F_0$ contours have not been studied in detail.

3. PSYCHOACOUSTIC EXPERIMENTS

3.1 Experiment 1

Experiment 1 clarifies whether speaker individuality still exists in the $F_0$ contours when spectral envelopes and amplitude contours are averaged for the three speakers, and when $F_0$ contours are modeled by the Fujisaki $F_0$ model.

3.1.1 Stimuli

The stimuli were original speech waves and speech waves re-synthesized by the LMA analysis-synthesis system. The synthesized speech wave is called an LMA speech wave. Four types of stimuli were used for the experiment 1:

(1-a) original speech waves,
(1-b) LMA speech waves without modification of their FFT cepstral data,
(1-c) LMA speech waves with spectral and amplitude envelopes averaged for the three speakers, which we call spectral-averaged LMA, and
(1-d) spectral-averaged LMA speech waves whose $F_0$ contours were modeled by Eq. (1), which we call $F_0$-modeled LMA.

The spectral-averaged LMA was calculated as follows. Since the time lengths of words uttered by the three speakers were different, a dynamic programming technique (DP) was adopted to shorten or lengthen each word non-linearly.

The local distance for the DP in this experiment was an LPC-cepstrum distance,

$$d(x, y) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (c_i^x - c_i^y)^2},$$

Fig. 4 Local DP-path constraint.

where $c_i^x$ and $c_i^y$ are $i$-th LPC-cepstra of speakers $x$ and $y$ and $p$ is the LPC order. The LPC-ceptra were analyzed using 16-th order LPC in 30 ms Hanning window at every 5 ms period. The local DP-path constraint is shown in Fig. 4.

The cepstral and amplitude sequences of each word were time-warped and their duration was normalized by the DP-path of each word. The duration-normalized FFT-cepstral and amplitude sequences were averaged arithmetically in each frame and inversely re-lengthened or re-shortened by each DP-path. The calculated FFT-cepstral sequence was the spectral-averaged FFT-cepstral sequence with the same length as the original. Thus, the spectral-averaged LMA (1-c) was re-synthesized with the spectral-averaged FFT-cepstral sequence, the averaged amplitude sequence and the extracted $F_0$ contour. In contrast, the $F_0$-modeled LMA (1-d) was re-synthesized with the spectral-averaged FFT-cepstral sequence, the averaged amplitude sequence and the $F_0$ contour modeled by Eq. (1).

The stimuli were presented through binaural earphones (STAX SRA-pro) at a comfortable loudness level in a sound-proof room (27.7 dB(A)). Each stimulus was presented to each subject six times in the experiment. Thus, the number of the presented stimuli in Experiment 1 was 216 (4 types x 3 words x 3 speakers x 6 times).

3.1.2 Subjects

The ten listeners (all males) serving as subjects were graduate students who were very familiar with characteristics of the speakers' voices. All listeners were native speakers of Japanese and had no known hearing impairments. They also served in the two other experiments discussed in 3.2 and 3.3.

3.1.3 Procedure

The task was to identify the speaker of the presented stimulus. When the subjects could not identify the speaker the first time, they were allowed to listen to the stimulus repeatedly. Speaker identification
rates for the stimuli were averaged for all subjects. This procedure was also used in the two other experiments discussed in 3.2 and 3.3.

3.1.4 Results and discussion

The speaker identification rates are shown in Fig. 5 and the F-test results with 1 and 18 free parameters are listed in Table 1. The significance level is $F(1, 18; 0.05) = 4.41$. The results lead to the following three conclusions:

1. Speaker individuality remains in the LMA analysis-synthesis speech. This is because $F(1, 18) = 2.82 < F(1, 18; 0.05) = 4.41$, i.e., the difference between the speaker identification rates of 99.1% for the original speech waves (1-a) and 97.4% for the LMA speech waves without modification (1-b) is not significant.

2. The speaker identification rate for the spectral-averaged LMA speech waves (1-c) is 92.0%, although $F(1, 18) = 7.35 > F(1, 18; 0.05) = 4.41$ between stimuli 1-b and 1-c). This indicates that there is speaker individuality in the $F_0$ contours, even though both spectral and amplitude envelopes are averaged, and that speaker individuality also exists in the spectral and amplitude envelopes, because the speaker identification rate for the spectral-averaged LMA speech waves (1-c) decreases by 5.4% from that of the LMA speech waves without modification (1-b). These results are consistent with previous research results in that spectral envelopes contain speaker individuality. Furthermore, the identification rates of 97.4% for (1-b) and 92.7% for (1-c) are still large enough to distinguish speakers.

3. Speaker individuality still remains in the $F_0$ contours calculated using Eq. (1), because the speaker identification rate for the $F_0$-modeled LMA speech waves (1-d) is 88.2% and $F(1, 18) = 1.67 < 4.41$ between stimuli (1-c) and (1-d). This result indicates that speaker individuality remains in the $F_0$-modeled LMA speech waves (1-d), although both the spectral and amplitude envelopes are averaged, meaning that the Fujisaki $F_0$ model plays an important role in describing the $F_0$ contours. This suggests that the Fujisaki $F_0$ model can be used as a base for controlling speaker individuality.

3.2 Experiment 2

Experiment 2 clarifies whether the three parameters $F_{0a}$, $A_{pa}$, and $A_{oa}$ of the Fujisaki $F_0$ model, which have large F ratio values, are effective in identifying speakers and whether perceived results of subjects are changed if these parameters are exchanged with those of other speakers.

3.2.1 Stimuli

The types of stimuli used in experiment 2 were:

(2-a) $F_0$-modeled LMA speech waves (same as (1-d)), and

(2-b) modified $F_0$-modeled LMA speech waves.

The modified $F_0$-modeled LMA speech waves (2-b) were re-synthesized with the spectral-averaged FFT-cepstral sequence and the modified $F_0$ contour, whose parameters $F_{0a}$, $A_{pa}$, and $A_{oa}$ were exchanged with those of another speaker. The duration of the modified $F_0$-modeled LMA speech waves was the same as that of the original speaker's speech waves. We call the speaker who contributes parameters $F_{0a}$, $A_{pa}$, and $A_{oa}$ for the stimuli for (2-b) the 'destination' speaker and the speaker who contributes parameters $\Delta T_{0a}$, $\Delta T_{pa}$, and $\Delta T_{oa}$ for (2-b) the 'origin' speaker. Figure 6 is a schematic diagram of the exchange of parameters.

Parameters were exchanged between speakers KI and KO, and YO's speech was used as dummy data. Thus, the number of the stimuli for (2-a) was 9, or 3 words $\times$ 3 speakers, and for (2-b) it was also 9, or 3 words $\times$ 2 speakers (parameters for KI and KO were exchanged) + 3 words $\times$ 1 speaker (YO speech was
used as dummy data). The stimuli for (2-a) were presented two times to the subjects first for training and the stimuli for (2-b) were presented six times randomly after presenting the stimuli for (2-a).

The procedures for this experiment were the same as those for experiment 1.

3.2.2 Results and discussion

The speaker identification rates for experiment 2 are shown in Fig. 7. The speaker identification rate of the $F_0$-modeled LMA (2-a) is the same as that of (1-d). The results suggest the following conclusions.

(1) The speaker identification rate of the destination speaker for modified $F_0$-modeled LMA speech waves (2-b-2) is 88.9% and $F(1, 18) = 0.05 \equiv F(1, 18; 0.05) = 4.41$ between (2-a) and (2-b-2) in Fig. 7. Note that the speaker identification rate of the $F_0$-modeled LMA speech waves (2-a) is 88.2%.

(2) The speaker identification rate of the origin speaker for modified $F_0$-modeled LMA speech waves (2-b-1) is 3.4%, and that for other speakers (2-b-3) is 7.8%. These values are much smaller than the rate identified for the destination speaker.

These results indicate that the parameters $F_0$, $A_p$, and $A_a$, which describe time-averaged $F_0$ frequency and a dynamic range of an $F_0$ contour, are significant in controlling speaker individuality. They also suggest that their timings are not particularly significant in identifying speakers, when the perceived speech is a word.
3.3 Experiment 3

Experiment 3 clarifies whether speaker individuality can be manipulated by shifting the time-averaged $F_0$ frequency. Time-averaged $F_0$ frequencies are often used for automatic speaker identification or verification. The experiment evaluates whether the parameter is also efficient for speaker individuality control.

3.3.1 Stimuli

The types of stimuli used in experiment 3 were: (3-a) spectral-averaged LMA speech waves (same as (1-c)) and (3-b) $F_0$-shifted LMA speech waves.

The $F_0$-shifted LMA speech waves were resynthesized with the spectral-averaged FFT-ceptral sequence and the $F_0$ contour whose time-average was shifted to equal that of another speaker. Since the time-averaged $F_0$ frequency can be modified by shifting $F_0$ contours, the Fujisaki $F_0$ model is not adopted in experiment 3. We call the speaker who contributes all parameters except the time-averaged $F_0$ frequency for the stimuli (3-b) the 'origin' speaker and the speaker who contributes the time-averaged $F_0$ frequency for (3-b) the 'destination' speaker.

$F_0$ contours were shifted between speakers KI and KO, and YO's speech was used as dummy data. Thus, the number of stimuli for (3-a) was 9, or 3 words $\times$ 3 speakers, and for (3-b) it was also 9, or 3 words $\times$ 2 speakers (KI and KO were exchanged) + 3 words $\times$ 1 speaker (YO speech was used as a dummy data). The stimuli for (3-a) were presented two times to the subjects first for training and the stimuli for (3-b) were presented six times randomly after presenting the stimuli for (3-a).

The procedures for this experiment were the same as those for experiment 1.

3.3.2 Results and Discussion

Table 2 F-test results of Experiment 3.

<table>
<thead>
<tr>
<th>Stimuli pair</th>
<th>(3-a)-(3-b-1)</th>
<th>(3-a)-(3-b-2)</th>
<th>(3-b-1)-(3-b-2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Result</td>
<td>24.63</td>
<td>67.50</td>
<td>2.59</td>
</tr>
</tbody>
</table>

![Fig. 8 Speaker identification rates of Experiment 3.](image)

(2) The results indicate no significant difference between the identification rate of the destination and the origin speakers for the $F_0$-shifted LMA speech waves, because $F(1, 18) = 2.59 < F(1, 18; 0.05) = 4.41$ between (3-b-1) and (3-b-2).

These results suggest that shifting the time-averaged $F_0$ frequency of one speaker to that of another speaker causes perceptual confusion during speaker identification. Although the time-averaged $F_0$ frequency still contains speaker individuality, these values are not as significant as the dynamics of the $F_0$ contours, which are described by the set of parameters $F_{0n}, A_p$, and $A_2$. Although it is not clear which is more significant, $F_0$ or $A_p$ and $A_2$, for speaker identification, it is clear that either one of them can cause misperception and both of them are needed to identify speakers.

4. CONCLUSIONS

This paper proposed some physical characteristics related to speaker individuality embedded in the $F_0$ contours of words and investigated the significance of the parameters describing the $F_0$ contours for speaker individuality control through psychoacoustic experiments. The Fujisaki $F_0$ model was used to extract and manipulate the physical characteris-
tics of $F_0$, and the LMA analysis-synthesis system was used to prepare synthesized stimuli with varied $F_0$ contours.

The results of Experiment 1 indicate that speaker individuality exists in both the $F_0$ contours and the spectral envelopes and speaker individuality remains in the $F_0$ contours calculated using Eq. (1). These results are consistent with previous research results.1-3)

Experiment 2 showed that the parameters $F_{0p}$, $A_{p}$, and $A_{o}$, which are relative to the dynamics of $F_0$ contours, are significant in identifying speakers. These results suggest that speaker individuality can be controlled when the three parameters are manipulated.

Experiment 3 showed that shifting the time-averaged $F_0$ frequency of one speaker to that of another does not provide a high speaker identification rate. Although the time-averaged $F_0$ frequency is often used as a distinctive feature for automatic speaker identification and verification, these values are not as significant as the dynamics of the $F_0$ contours illustrated using the parameters $F_{0p}$, $A_{p}$, and $A_{o}$.

The three experiments were completed for about three months. Although the speaker identification rates rose a little in the period, the difference of the speaker identification rates through the period is not significant. Thus, the experiments are consistent.

The speech data used in these experiments were three-mora words with accent second mora, such as “aōi” (blue). Additionally, the speakers for the speech data come from different districts. In future work, we will investigate speaker individuality in different accent or mora words, in sentences, and in the same dialect speakers. Duration of phonemes and pauses may also become significant in identifying speakers when listening to sentences and relative significance of the three parameters $F_{0p}$, $A_{p}$, and $A_{o}$ may change when using the same dialect speech data.

ACKNOWLEDGEMENTS

This work was supported by Grant-in-Aid for Scientific Research from the Ministry of Education (No. 07680388).

REFERENCES


Masato Akagi was born in Okayama, Japan on September 12, 1956. He received a B.E. degree from Nagoya Institute of Technology in 1979, and M. E. and D.E. degrees from Tokyo Institute of Technology in 1981 and 1984, respectively. He joined the Electrical Communication Laboratories, Nippon Telegraph and Telephone Corporation (NTT) in 1984. From 1986 to 1990, he worked at the ATR Auditory and Visual Perception Research Laboratories. Since 1992, he has been with School of Information Science, Japan Advanced Institute of Science and Technology, Hokuriku (JAIST) and now he is an Associate Professor of JAIST. His research interests include speech perception, modeling of speech perception mechanisms of humans, and signal processing of speech. During 1988, he joined the Research Laboratories of Electronics, MIT as a visiting researcher and in 1993, he studied at the Institute of Phonetic Science, Univ. of Amsterdam. Dr. Akagi is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan, the Acoustical Society of Japan (ASJ), the Institute of Electrical and Electronic Engineering (IEEE), the Acoustical Society of America (ASA), and the European Speech Communication Association (ESCA).

Taro Ienaga was born in Kanagawa, Japan, on May 1, 1970. He received a B.S. degree in Physics from Science University of Tokyo in 1993, and the M.E. in Information Science from Japan Advanced Institute of Science and Technology (JAIST) in 1995. He is currently a software engineer of Sony Corporation. He is a member of the Acoustical Society of Japan, and the Institute of Electronics, Information and Communication Engineers of Japan.

80