Title	文音声中に含まれる個人性情報の知覚に関する研究
Author(s)	鈴木，教郎
Citation	
Issue Date	1999-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1258
Rights	
Description	Supervisor:赤木 正人，情報科学研究科，修士

A study on perception of speaker individuality embedded in sentence utterance

Norio Suzuki

School of Information Science,
Japan Advanced Institute of Science and Technology

February 15, 1999

Keywords: speaker individuality, STRAIGHT, temporal decomposition.

1 Introduction

The aim of this work is to clarify perception of speaker individuality embedded in sentence utterances. Human uses speaker individuality. Physical correlate corresponding to speaker individuality has not been clarify yet.

If physical cowelates of speaker individuality are clarified, they can be applied to varies speech processing techniques as : (1)text-to-speech synthesis and (2) speech morphing.

Ienaga and Akagi have shown that there is speaker individualities in fundamental frequency(F0) contours, Kitamura, Akagi and Kitazawa have investigated perceptual effect of dynamic features of continuous vowel for speaker identification.

These studies, however, have not clarified a speaker individuality in the spectral trajectory and the F0 contour at the same time. It is necessary to verify whether speaker individuality is related to spectra, F0 and dynamics.

This paper investigates relations between speaker individuality embedded in sentence utterance and its physical correlates. Spectra, F0 and their dynamics in sentence utterance were analyzed by STRAIGHT and S^{2}BEL-TD and psychoacoustic experiments.

2 STRAIGHT

The stimuli used for the experiments is resynthesized by STRAIGHT analysis-synthesis system[3] on the condition shown in Table 1.

The spectra were converted into LSF(Line spectrum frequency) parameters, because interpolation characteristics of LSF are excellent.

Table 1: Analysis-synthesis condition

frame length	40ms
frame shift	1ms

3 S^{2}BEL-TD

To expose dynamics of spectra and fundamental frequencies(F0), the spectral dynamics in sentence utterances was analyzed by S^{2}BEL-TD as follows.

- S^{2}BEL-TD involves the decomposition of spectral parameters into a sequence of overlapping event functions and an associated sequence of event targets, as shown in Eq.(1).

$$\hat{\mathbf{y}}(n) = \sum_{k=1}^{K} \mathbf{a}_k \phi_k(n), \quad 1 \leq n \leq N \tag{1}$$

where,$\mathbf{a}_k$ and $\phi_k(n)$ are the k^{th} event target, i.e.spectral target, and the k^{th} event function, respectively. $\hat{\mathbf{y}}(n)$ is the approximation of $\mathbf{y}(n)$.

Event function $\phi_k(n)$ means a rate of before and after spectra.

Similarly, the F0 and gain parameters analyzed by STRAIGHT and can be reconstructed using the F0 targets, gain targets,$p(n)$ and $g(n)$, are event functions$\phi_k(n)$ as follows.

$$\hat{p}(n) = \sum_{k=1}^{K} p_k \phi_k(n), \quad 1 \leq n \leq N \tag{2}$$

$$\hat{g}(n) = \sum_{k=1}^{K} g_k \phi_k(n), \quad 1 \leq n \leq N \tag{3}$$

Where, $\hat{p}(n)$, $\hat{g}(n)$, p_k and g_k are the reconstructed F0 parameter, gain parameter for the n^{th} frame, the k^{th} pitch target and the k^{th} gain target respectively.

Number of events is regarded to be the same as that of phoneme, to compare event values among speakers.

4 Analysis of difference in fundamental frequencies(F0), spectra and their dynamics

in this section, distances between F0s, spectra and dynamics extracted by S^{2}BEL-TD are analyzed.

speech data Speech data for all the experiments are sentences such as ”いいえ, うえにある” uttered by five male speakers.

The sentences were recorded at a sampling rate of 48 kHz. Those sentences were down sampled at a sampling rate of 8 kHz.

parameters Parameters extracted by S^{2}BEL-TD were spectra, F0 and their dynamics.

method Distances of the parameters extracted by the S^{2}BEL-TD between speakers are calculated.

the spectral parameter LSF is converted into LPC and LPC is converted into cepstrum. Cepstrum distortion $\mathbf{cd}_k$ and the mean spectral distortion **CD** were

$$\mathbf{cd}_k = D_b\sqrt{2\sum\nolimits_{i=1}^{p}(c_{ik}^{(x)} - c_{ik}^{(y)})^2} \tag{4}$$

$$\mathbf{CD} = \frac{1}{K}\sum\nolimits_{k=1}^{K}\mathbf{cd}_k \quad , \tag{5}$$

where p is cepstrum order, K is the number of events and $c_{ik}^{(x)} and c_{ik}^{(y)}$ are k^{th} ceptrum parameters of speaker x and y, respectively. D_b is $D_b = 10/\ln 10$.

Similarly, Eq(6) represents the distortion $\mathbf{DIST_{F0}}$ of F0 and the distortion $\mathbf{DIST_g}$.

$$\mathbf{DIST_{F0}} = D_b\sqrt{\sum\nolimits_{i=1}^{K}(p_i^{(x)} - p_i^{(y)})^2} \tag{6}$$

$$\mathbf{DIST_g} = D_b\sqrt{\sum\nolimits_{i=1}^{K}(g_i^{(x)} - g_i^{(y)})^2} \quad , \tag{7}$$

where $p_{ik}^{(x)} and p_{ik}^{(y)}$ are k^{th} F0 parameters of speaker x and y. $g_{ik}^{(x)} and g_{ik}^{(y)}$ are k^{th} gain parameters of speaker x and y.

The distortion $\mathbf{DIST_t}$ for dynamics is length on sentences.

$$\mathbf{DIST_t} = D_b\sqrt{\sum\nolimits_{i=1}^{K}(s_i^{(x)} - s_i^{(y)})^2}, \tag{8}$$

where $s_{ik}^{(x)} and s_{ik}^{(y)}$ are k^{th} phone lengths of speaker x and y.

results Fig.1 shows plots of placement according to the distortions of spectra and F0s. Fig.2 shows plots of placement according to the distortions of gains and phoneme lengths.

Placement according to spectra shows that speaker a and speaker e are father away from other one, where as that speaker b and speaker d were same place.

Placement according to F0 shows that speaker a is father away from the mean, where as speaker b and speaker d were same place.

Placement according to gain shows that speaker c is near to mean.

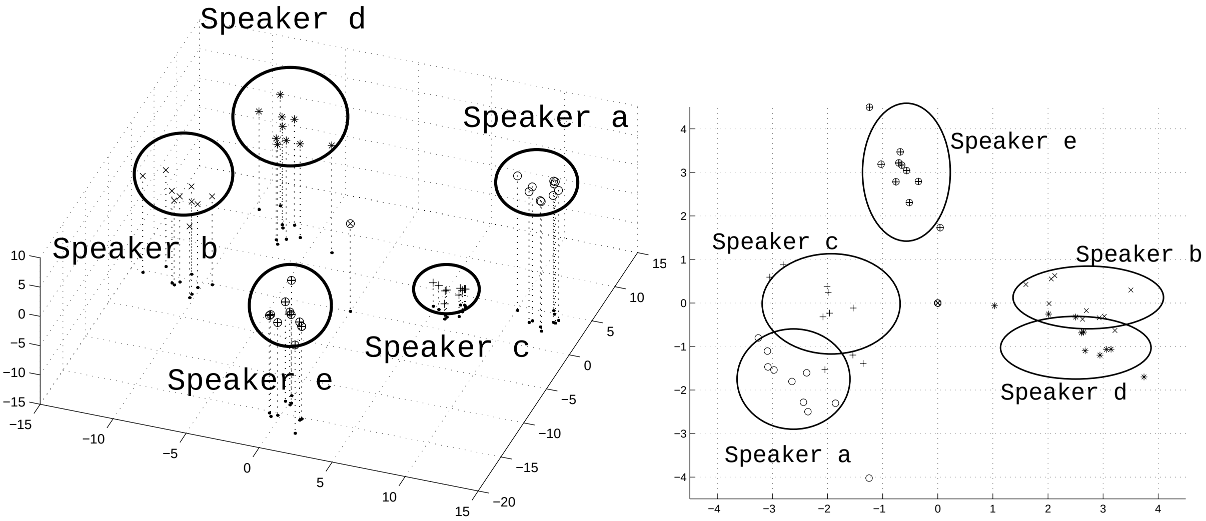

Figure 1: placement according to the distortion(left: spectrum, right: fundamental frequency)(speaker a: $\circ$, speaker b: $\times$, speaker c: $+$, speaker d: $*$, speaker e: $\oplus$, average: $\otimes$)

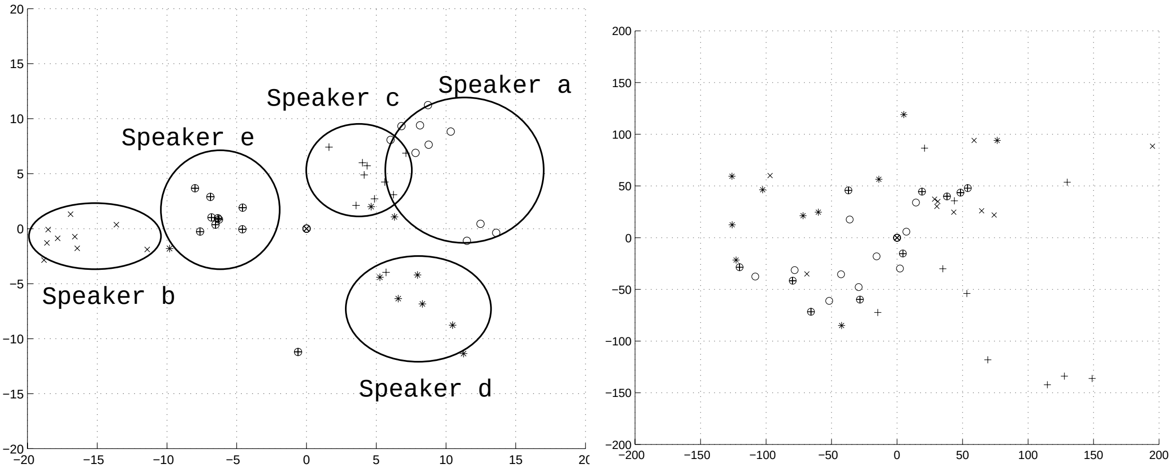

Figure 2: placement according to the distortion(left: gain, right: phoneme)(speaker a: $\circ$, speaker b: $\times$, speaker c: $+$, speaker d: $*$, speaker e: $\oplus$, average: $\otimes$)

5 Experiment

The relationship between physical characteristics and the speaker identification rates was studied by using stimuli in which several types of physical characteristics were varied.

5.1 Experiment1: existence of speaker individualities

Experiment 1 shows the existence of speaker individualities in the synthesis speech.

speech data Speech data was the same as analyzed data. This experimental speech data is also used in Experiment 2 and Experiment 3.

stimuli In the experiment, the following four types of stimuli were used:

1-A. original speech waves,
1-B. STRAIGHT analyzed-synthesized speech waves,
1-C. spectra of 1-B are converted to into LSF,
1-D. three parameter were converted into event targets and event functions by S^2BEL-TD.

subjects Ten listeners(ten males) serving as subjects in Experiment 1, 2 and 3 were graduate students who were very familiar with speaker voice characteristics. All were native speakers of Japanese.

procedure The stimuli were presented through binaural earphones at a comfortable loudness level. The task was to identify speakers. Method was Naming method. This experimental procedure is also used in Experiment 2 and 3. Subject listened to stimuli 1-A 30 times (6 times per one speaker), and listened to other stimuli 45 times (9 times per one speaker).

result and discussion The identification rates for Experiment 1 are shown in Fig.3.

Speaker identification rates on stimuli 1-B was 99.8% where as , it is 100 % for other stimuli 100%.Speaker identification rates were evaluated using the F-test. There speaker individualities in the stimuli 1-B(Fig.3). Thus, there are speaker individualities in these stimuli.

5.2 Experiment2: Speaker individuarities by each parameters

Experiment 2 shows speaker individualities in the spectra, F0 and gain. Dynamics is assumed to have speaker individualities.

stimuli Stimuli were synthesized data, whose parameters were averaged. The stimuli are as follow:

2-A. same as stimuli 1-D,
2-B. gains of stimuli 1-D are averaged among speakers,

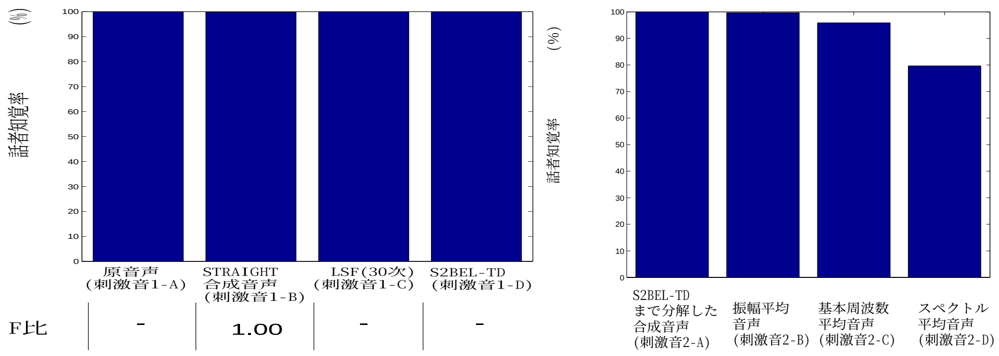

Figure 3: speaker identification rate{exp1(left)},{exp2(right)}{F(1,18:0.05)=4.41}

2-C. F0s of stimuli 1-D are averaged among speakers,
2-D. spectra of stimuli 1-D are averaged among speakers.

procedure This experimental procedure is the same as Experiment 1. Subjects listened to 45 stimuli (9 times per one speaker).

results and discussion Speaker identification rates for Experiment 2 are shown in Fig.3.

Speaker identification rates are 100 %, 99.6%, 95.8% and 79.6 % for stimuli 2-A, 2-B, 2-C and 2-D, respectively.

Table 2: F-test

	stimuli 2-B	stimuli 2-C	stimuli 2-D
F-rate(for stimuli 2-A)	1.00	29.8	43.4
F-rate(for stimuli 2-B)	-	17.8	40.7

{F(1,18:0.05)=4.41,F(1,18:0.01)=8.28}

Speaker identification rates were evaluatied by using the F-test with 1 and 18 free parameters. The results lead to the following two conclusions.

1. There are no speaker individualities in the gain(Table 2).
2. There are speaker individualities in the spectra(Perception rate of 2-D is 79.6%).

5.3 Experiment3: Dynamics and Perception

Experiment3 was performed to investigate significant physical cues in dynamics, spectra and F0 for speaker identification.

Table 3: Psychoacoustic Experiment

Speaker	5
Subject	10
Headphone	SENNHEISER HDA 200
Headphone Amp	SANSUI AUα-907MR
Hearing level	75〜80 dB (A)

Table 4: Perception result

		identification rate(%)
exp1(stumili 3-A)	all same	100
exp2(stumili 3-B)	LSF and F0	98.8
	LSF and dynamics	83.8
	F0 and dynamics	20.3
exp3(stumili 3-C)	LSF	71.1
	F0	11.1
	dynamics	7.0

stimuli Three values were exchanged between speakers. The stimuli are as follow:

3-A. three parameters come from the same speaker (5times)
3-B. two parameters come from the same speaker (60times)
3-C. each parameter comes from each speaker (60times)

results and disccusion The experiment results are shown in Table 4. Table 4 shows that the result of stimuli 3-A 3-B and 3-C.

Table 5, Table 6 shows details of identification rates of spectra and dynamics, and fundamental frequency and dynamics.

Table 7 shows details of identification rates of stimuli 3-C.

The experimental results and the analysis results inducate that (1) when all parameters are same subjects can perceive perfectly , (2) there are much speaker individuality in the static components of spectra(Exp.1,2 and 3) , (3) the connection of static components of spectra and fundamental frequencies give high speaker identification rates , (4) the speaker can be identified even though only the static components of spectra come from the speaker, when the spectral distance between one speaker and others is large , (5)speaker individuality can be controlled by manipulating spectra and fundamental frequencies.

The experimental results show the same results as the report [5], that is, speaker individuarity is affected by amount of differences between acoustic features.

Table 5: Identification rate(LSF and dynamics of stimuri 3-B)

			speaker			
	a	b	c	d	e	other
speaker a	95.0*	0	0	1.7	1.7	1.7
speaker b	0	78.3*	5.8	6.7	0.8	8.3
speaker c	0	3.3	84.2*	1.7	0	10.8
speaker d	0	5.0	5.8	67.5*	0.8	20.8
speaker e	0	0	1.7	1.7	94.2*	2.5

Table 6: Identification rate(F0 and dynamics of stimuli 3-B)

			speaer			
	a	b	c	d	e	other
speaker a	0.8*	12.5	25.0	5.8	24.2	31.7
speaker b	23.3	29.2*	14.2	7.5	24.2	1.7
speaker c	25.0	12.5	25.0*	9.2	22.5	5.8
speaker d	23.3	3.3	8.3	40.0*	22.5	2.5
speaker e	23.3	19.2	22.5	20.8	6.7*	7.5

6 Conclusion

In order to investigate speaker individuality in spectra, fundamental frequency and their dynamics, acoustic feature extraction by STRAIGHT and S^{2}BEL-TD, analysis of differences between featuers, and the psychoacoustic experiments were carried out.

The results indicate that (1) there are much speaker individuality in the static components of spectra , (2) the connection of static components of spectra and fundamental frequencies give high speaker identification rates and (3) the speaker can be identified, even though only the static components of spectra come from the speaker, when the spectral distance between one speaker and others is large.

Table 7: Identification rate(all differ of stimuli 3-C)

			speaker			
	a	b	c	d	e	other
speaker a	95.6*	0	0	2.8	1.1	0
speaker b	0	62.5*	21.1	13.9	2.5	7.5
speaker c	6.1	6.9	76.9*	9.7	0	5.0
speaker d	0	28.6	25.6	44.7*	1.1	19.7
speaker e	0	1.1	2.8	0	95.6*	1.7

References

[1] T.Ienaga and M.Akagi "Speaker individuality in fundamental frequency contours and its control" J. Acoust. Soc. Jpn.(E), **18**, 2, 1997

[2] T. Kitamura, M. Akagi and S. Kitazawa "Perceptual effect of spectral trajectory patterns for speaker identification" ASJ Tech. Report, H98-97, 1998

[3] H. Kawahara, I. Masuda-Katsuse and K. Toyama Compensatory time window for speech analysis, modification and synthesis using STRAIGHT" ASJ Tech. Report, H97-47, 1997

[4] A. C. R. Nandasena and M. Akagi "Spectral stability based event localizing temporal decomposition" Proc. ICASSP98, II, 957-960

[5] M. Hashimoto, S. Kitagawa and N. Higuchi "Quantitative analysis of acoustic features affecting speaker identification" J.Acoust. Soc. Jpn, Vol54, No3, 1998