A study on speech morphing based on the exchanging events about spectra and fundamental frequencies

Fujino Yoshiyuki

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

February 15, 2001

Keywords: speech morphing, subset data, vowel spectra .

1 Introduction

"Morphing" is a well known technique in image processing, that gradually changes the object to another one. Speech morphing produces similar results for speech. There are many concepts in speech morphing. In this paper, speech morphing is defined as the speech character transformation from one person's speech to that of someone else from the point of view of speaker individuality.

Although most speech morphing system can synthesize speech, they still cannot synthesize speech with various voice quality such as speaker individualities; In order to control speaker individualities, therefore, they need a large database about various voice characteristics(fullsize data) for both original speech and target speech, and also need the complicated processing[1].

This paper proposes a new approach to speech morphing by using less voice characteristics(subset data) for the target speaker, assuming that there exist the fullsize data for the original speaker. Additionally infuluence of morphed speech by morphing parameters is investigated.

2 Speech analysis-synthesis system

In proposed method of speech morphing vowel spectra and fundamental frequencies are principally used as morphed parameters for subset data. To obtain these parameter original speech is analyzed by STRAIGHT analysis-synthesis system[2] to analyze spectra and

fundamental frequencies. The spectra are converted into LSF(Line spectral frequency), because interpolation characteristics of LSF are excellent. The spectra is, furthermore, analyzed by S^2BEL-TD[3], which decomposes spectra into the consonant segment and vowel segment.

3 Morphing Parameters

One of the most important points to perform the speech morphing by using the less amount of subset data is the selection that what kind of morphing parameters should be. In this section, some morphing parameters to perform the speech morphing are introduced.

3.1 Speech data for morphing

3.1.1 Speaker A (fullsize data)

A male speaker MMS from ATR Japanese speech database is used as speaker A. A word "そびえる" which compose of vowel and some voiced-unvoiced consonants is used as speech data.

3.1.2 Speaker B (subset data)

For obtaining the subset data for speech morphing, a male speaker who is the postgraduate student is used as speaker B.

3.2 Isolated vowel event

Sentences are composed by words, which are composed by syllables. Japanese syllables are usually composed by pair of vowel and consonant. Furthermore, duration of vowel is longer than that of consonant and spectra of vowels are also clearer than those of consonants. So vowel is usually recognized easily and certainly. This is the reason why vowel is often used in speech recognition by both human and machines.

For the least morphing parameters for subset data, 5 isolated vowels (/a/, /i/, /u/, /e/, /o/) which uttered by the target speaker B is used in speech morphing.

3.3 Nutralization

Coarticulation is produced in continious speech, which make difference the acoustic features from that of isolated speech. Because of that, nutralization is yielded by coarticulation. To give nutralization in speech synthesizing the more natural morphed speech is synthesized.

The second vowel event in 3 connected on concatenated vowels is used as the morphed parameters for speech morphing with nutralization.

3.3.1 Speech data (3 connected on concatenated vowels)

The morphing parameters to perform speech morphing is 3 continious vowel /oie/ and /ieu/ uttered by the target speaker B. These are corresponded to the speech data "そびえる /sobieru/". The second vowel event /oie/, /iiu/ is used. Particulaly /oie/ and /ieu/ are used as the first vowel /o/ and last vowel /u/ in speech data "そびえる", respectively.

3.4 Fundamental frequency (F0s)

There are much speaker individuality in dynamics of F0 contours[?]. In this paper speaker individuality in relation to F0s are used as morphing parameters.

3.4.1 The mean value of F0s

The mean value of F0s is used as the least subset data in relation to F0s.

3.4.2 F0s event

Dynamics are differ between the speakers respectively. The timbre is changed by the difference of dynamics of F0 contours, i.e. the magnitude of accent.

So the typical of speech data about accent type is used in subset data assuming that the magnitude of accent in words are same in same speaker.

4 Experiment

Listening experiments by ABX-test are done to confirm however the morphed speech are morphed to the target speaker, and to investigate influencies for the morphed speech by using morphing parameters.

Table.1 shows how to perform speech morphing, Table.2 shows the list of morphed speech used in experiments.

Experiment results by using ABX-test are shown in Fig.1, in which -3 in horizontal axis means near the speaker A, on the other hand, 3 near the speaker B. Resulted speech were evaluated using F-test and T-test. Mean values of speech by X-1 are different to those of by X-2 from Figure.1. But it is evaluated by statistic test that speech by X-1 is not almost different to those of by X-2.

It was said that the morphed speech by using the second vowel event in 3 connected on concatenated vowels is little percepted to be near in the vowel of target speaker than that of by using isolated vowel events; nevertheless the former generates much smooth and natural speech on its quality than latter. Because of that, it was confirmed the nutralization by coarticulation can represent the naturality of the synthesized speech.

The morphing parameters which yield to morphed speech are vowel event, consequently.

The change of voice quality by using F0s is not more conspicuous than that of spectra.

Table 1: method of speech morphing

operate	content
Fev	exchange the F0s' event of speaker A to the target speaker B
Fav	exchange the mean of F0s of speaker A to the target speaker B
X-1	exchange the vowel event of speaker A to isolated vowel event of target speaker B
X-2	exchange the vowel event of speaker A to the second vowel event in 3 connected on concatenated vowels of target speaker B
X-3	exchange the vowel event of speaker B to the second vowel event in 3 connected on concatenated vowels of target speaker B

Table 2: morphed speech in experiments

音声	a	b	c	d	e	f	g	h	i
Fev	○				○			○	
Fav		○		○			○		
X-1			○	○	○				
X-2						○	○	○	
X-3									○

The results yielded that vowel spectra, mean of F0s and F0s' event are the most effective morphing parameters to perform speech morphing by using subset data.

5 Conclusion

Speech morphing by using subset data is performed. Also, the morphing prameters to perform the effective speech morphing are defined in this research.

The results show that 3 connected on concatenated vowels event, mean of F0s and F0s' event are the morphing parameters in order to perform the effective speech morphing. It is to say, sentences which are composed by most the 3 combination continious vowel and by the most typical words which include for accent informations is need for subset data in order to perform speech morphing.

The resulted speech are not representated speaker individuality of the target sufficiently, however, in term of overall performance, the proposed speech morphing by usin subset data efficientlly generates smooth speech.

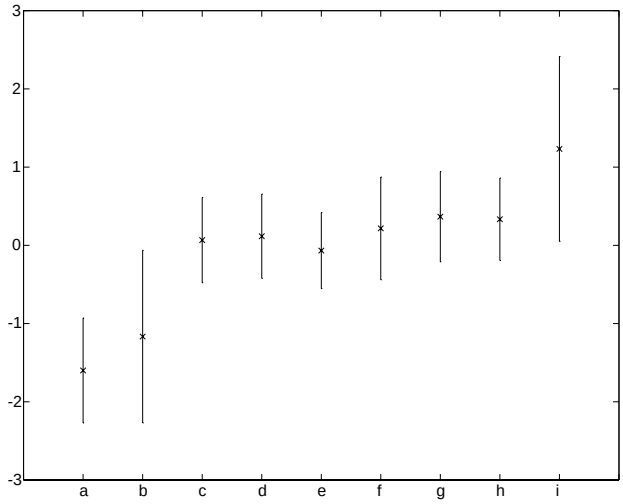

Figure 1: placement, mean and the standard deviation for morphed speech “そびえる”

References

[1] Masanobu ABE, “Speech morphing by gradually changing spectrum parameter and fundamental frequency,” IEICE SP96-40, June, 18-19, 1996.

[2] H,Kawahara, I.masuda-Katsuse, K.toyama, “ACompensatory time widow for speech analysis, modification and synthesis using STRAIGHT,” ASJ Tech. Report, H97-47,1997.

[3] A.C.R.Nandasena and M.Akagi, “Spectral Stability Based Event Localizing Temporal Decomposition,” Proc.ICASSP98, II, 957-960

[4] Masato Akagi and Taro Ienaga, “Speaker individuality in fundamental frequency contours and its control,” J. Acoust. Soc. Jpn. (E) 18, 2(1997)