SPEECH MORPHING BY GRADUALLY CHANGING SPECTRUM PARAMETER AND FUNDAMENTAL FREQUENCY

Masanobu Abe

NTT Human Interface Laboratories
1-2356 Take, Yokosuka-Shi, Kanagawa, 238-03 Japan

ABSTRACT

This paper proposes a new application of speech modification called "speech morphing". In image processing, morphing is a well known technique that gradually changes one person's face to that of someone else. Speech morphing produces similar results for speech; i.e., one person's speech is gradually changed to that of someone else. Speech morphing makes it possible to create movies or multi-media entertainment together with image morphing. The proposed algorithm pitch-synchronously modifies fundamental frequency(F_0) and DFT spectrum and outputs high quality speech. To clarify the balance of F_0 modification and spectrum modification, listening tests were carried out using 20 male speakers. The results yielded the relationship between the amount of modification and speaker identity. In terms of overall performance, listening tests show that the proposed algorithm successfully generates smooth, high quality voice changes.

1. INTRODUCTION

In this paper, we propose a new application of speech modification, we name it speech morphing. In image processing, morphing is a well known technique that gradually changes, for example, one person's face to that of someone else, in an extreme case it might become an animal's face. Speech morphing produces similar results for speech; i.e., speech from speaker A is gradually changed to match that from speaker B. Our aim for speech morphing is to produce a smooth change in speech identity; the change is to be perceived by the listener as it occurs.

Speech morphing might have several applications. One example is to provide a new tool to create movies or multimedia entertainment together with image morphing. Another example is to simplify the development of speech message systems. In the system, where pieces of pre-recorded speech are concatenated to generate new messages, speech quality must be kept constant regardless of the speaker, recording conditions and so on. However, this requirement is not always satisfied. In such cases, speech morphing makes it possible to smoothly concatenate segments of different quality. Section 2 introduces the proposed algorithm. In section 3, listening tests are carried out to determine the importance of F_0 and spectrum modification in changing speech identity, and overall performance is evaluated.

2. A SPEECH MORPHING ALGORITHM

Inputs of the speech morphing algorithm are speech uttered by speaker A and speaker B, and they are assumed to contain the same phoneme sequences. Output consists of speaker A's speech, modified speech, and speaker B's speech. By temporally controlling speech parameters, the identity of the modified speech gradual changes. The control parameters are fundamental frequency(F_0) and speech spectrum.

2.1. Outline of the algorithm

Figure 1 shows a block diagram of the proposed algorithm. At this stage, the proposed algorithm only modifies voiced segments. In the following explanation, the numbers refer to the block numbers cited in Fig. 1.

(1) Assign phoneme boundaries to both speaker A's and speaker B's speech.

(2) Assign pitch marks [1] to both speaker A's and speaker B's speech.

(3) For voiced phoneme segments, find pitch mark correspondence between speaker A's and speaker B's speech, as shown in Fig. 2.

Steps (4), (5), (6), and (7) are performed pitch-synchronously.

(4) Set values for the amount of modification desired. The values change with time.

(5) According to the output of step (4), modify the speech spectrum. Details are explained in the following section.

(6) According to output of step (4), modify F_0 using the TD-PSOLA algorithm [1].

(7) Synthesize speech by pitch-synchronous overlap addition.

2.2. Spectrum parameter modification algorithm

Figure 3 shows procedures for spectrum parameter modification. For each pair of corresponding pitch marks determined in step (3), by setting pitch mark position as the center of an analysis window, speech waveforms are extracted from speaker A's and speaker B's speech. The window length is twice as long as the shorter pitch period of the pair of pitch marks. Spectra of the waveforms are calculated by Fast Fourier Transform(FFT). In the spectrum domain, a new set

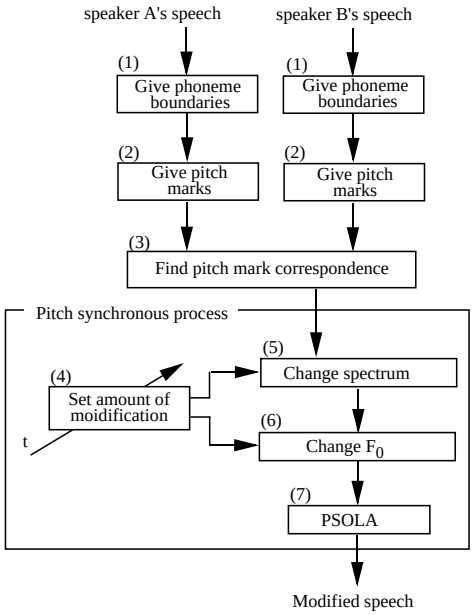


Fig.1 A block diagram of a speech morphing algorithm

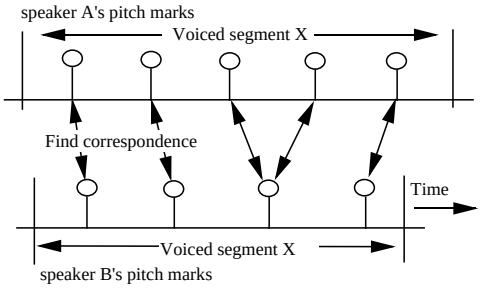


Fig. 2 Find pitch mark correspondence

of FFT coefficients are generated by mixing FFT coefficients of speaker A and speaker B; i.e., FFT coefficients below and above α Hz are copied from speaker A's and speaker B's FFT coefficients, respectively. The threshold α Hz is set in step (4). Finally, by inverse Fast Fourier Transform, two-pitch-length waveforms are obtained.

3. PERFORMANCE EVALUATION

The aim of speech morphing is to change speech identity smoothly between two speakers. Therefore, we have to know the relationship between the amount of acoustic modification and the amount of psychological difference caused. To investigate this point, listening tests were carried out in terms of spectrum modification, F_0 modification, and simultaneous modification of both spectrum and F_0. Finally, we evaluate the overall performance of speech morphing. Steps (1) and (2) were performed semi-automatically, and errors were manually corrected.

3.1. Speech data

In the following experiments, we used the same database created by 20 male speakers uttering the same set of 520 words; the total number of utterance was 10,400. From all speaker combinations (190 pairs), we selected speaker pairs

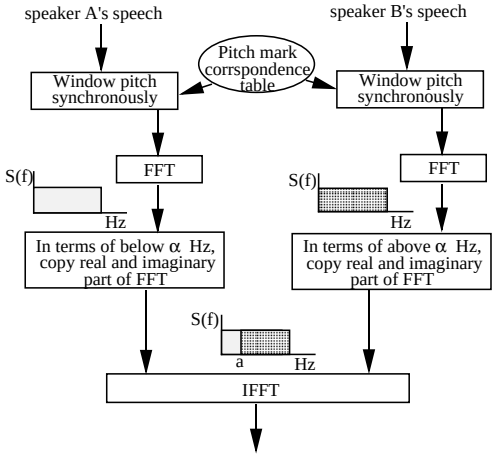


Fig. 3 A block diagram of spectrum modification

Table 1 Analysis conditions

sampling frequency	12kHz
FFT	512 points
window length	21.3msec
cepstrum order	20
threshold (α)	300Hz(3bark) 642Hz(6bark) 1214Hz(10bark) 2360Hz(14bark) 3892Hz(17bark)

for the listening tests. Details of the selection process are explained in each experiments.

3.2. Effect of spectrum modification

To measure the degree of spectum mixing, the following values were calculated.

$$R_{spec} = \frac{\int_0^{\alpha} S_{syn}(f)df}{\int_0^{\beta} S_{syn}(f)df}$$

$$R_{err} = \frac{\int_0^{\alpha} |S_{spA}(f) - S_{spB}(f)|df}{\int_0^{\beta} |S_{spA}(f) - S_{spB}(f)|df}$$

R_{spec} and R_{err} involve the ratio of the two speakers' power spectrum and of their differences, respectively. Here, S_{syn} is the average power spectrum envelope of synthesized speech. S_{spA} and S_{spB} are the average power spectrum envelopes of speaker A and B, respectively. α is the threshold value explained in section 2.2, and β is Nyquist frequency. The power spectrum envelope is obtained from FFT cepstrum analysis. Analysis conditions are shown Table 1. Figure 4 plots the R_{spec} and R_{err} values for the four speaker pairs that yielded the maximum and minimum values on average. It is clearly shown that R_{err} values depend on the distance between the speaker pair. The following experiment used the speech data used to plot the R_{err} curves in Fig. 4.

ABX listening tests were carried out to check how spectrum mixing impacts speaker identification. Stimuli A and B were speech uttered by speaker A or speaker B, and stimulus X was synthesized speech by mixing the spectra of speaker A and speaker B. The threshold values for the mixed spectrum are shown in Table 1; they were not changed over time, and fixed so as to synthesize a stimulus. Listeners were asked

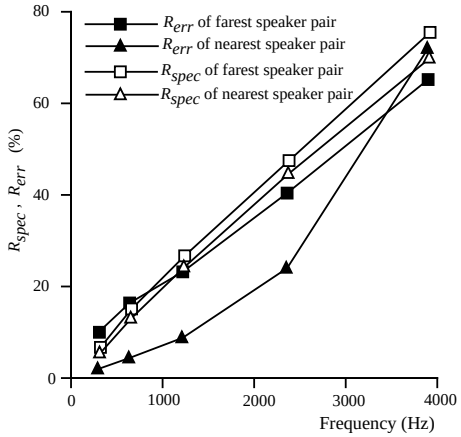


Fig. 4 Spectrum differences between speakers

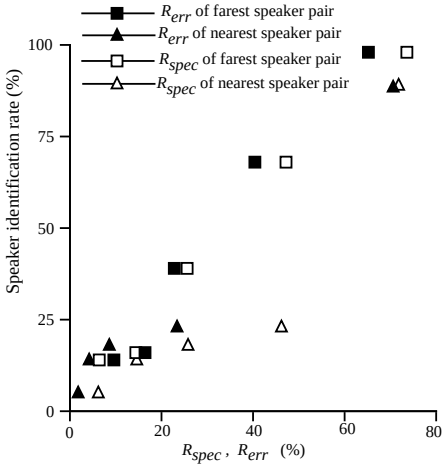


Fig. 5 Impact of spectrum mixing on speaker identification

to select either A or B as being closest to X. In order to eliminate unwanted cues, all stimuli were synthesized using speaker A's F_0 contour and duration. The ABX triads were synthesized using word utterance, and the total number of triads was 28. Eleven listeners participated.

Figure 5 shows the relationship between R_{spec}, R_{err} and the speaker identification rate; the rate at which a stimulus whose spectrum (below α Hz) copied from a speaker was judged as coming from that speaker. Because the speaker identification rate correlates strongly with R_{err} in Fig. 5, R_{err} is a good measure for estimating the impact of spectrum mixing. If α is set so that R_{err} becomes 40%, the output speech is judged as different from that of the original speaker.

3.3. Effect of F_0 modification

To investigate how F_0 modification impacts speaker identification, opinion tests were carried out. The average F_0 difference between speaker A and speaker B was 22Hz, and speech was synthesized in 5 F_0 ranges within the F_0 difference. Here, again, in order to eliminated unwanted cues, speech samples were synthesized using speaker B's spectrum and duration. Stimuli consisted of two speech samples in each F_0 range. The same eleven listeners were asked to rate

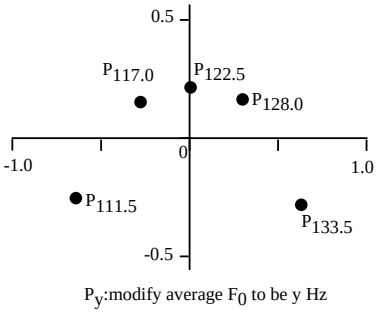


Fig. 6 Effect of F_0 modification

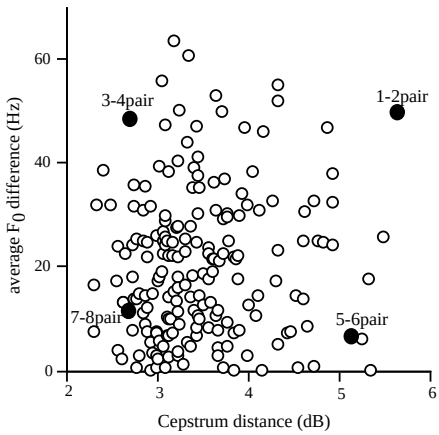


Fig. 7 Acoustic distance between speakers

the similarity of each pair into five categories ranging from "similar" to "dissimilar". The total number of stimuli was 20. Hayashi's fourth method of quantification [2] was applied to the experimental data obtained by the listening test. This method places a sample in a space according to the similarities between the samples.

The projection of the results onto a two-dimensional space is shown in Fig. 6. This figure shows the relative similarity-distance between stimuli. It is observed that F_0 range differences were clearly distinguished by the listeners, and that there is a linear relationship between F_0 range difference and psychological difference. Therefore, it is expected that if F_0 range is linearly increased or decreased over time, listeners hear linear changes over time.

3.4. Spectrum modification vs. F_0 modification

The effects of spectrum modification and F_0 modification depend on how large the spectrum and F_0 range differences are between speakers; i.e., if there is a large difference in spectrum, spectrum modification is more important than F_0 modification, and vise versa. We investigated this point. Figure 7 shows cepstrum distance and F_0 range difference for all speaker combinations. In the following experiments, speech from pairs 1-2, 3-4, 5-6, and 7-8 were used.

In terms of similarity, pair comparison tests were carried out. Each speech pair consisted of a word utterance from 4 different groups. The groups were original speaker's speech, speech whose F_0 or spectrum or both F_0 and spectrum were modified from original speaker to target speaker. Here, we assume that the fourth group approximates the

target speaker's speech. The 11 listeners were asked to rate the similarity of each pair into five categories ranging from "similar" and "dissimilar". Again, Hayashi's fourth method of quantification was applied to the experimental data obtained by the listening test.

Figures 8 and 9 show the relative similarity-distance between stimuli. As shown in Fig. 8, for pair 3-4, F_0 modification has a large enough impact to change speaker identity because F_0 range difference is large while the spectrum difference is only small. On the other hand, for pair 5-6, the spectrum modification was not large enough to change speaker identity. If we use only spectrum modification, the speakers must have much more different spectra than pair 5-6 in order for the speaker identity change to clearly heard.

3.5. Evaluation of overall performance

To evaluate overall performance, listening tests were carried out. Three pairs were selected based on cepstrum differences, F_0 range differences, and F_0 contour differences. Using short sentences, 4 kinds of stimuli were prepared; i.e., natural speech, morphed speech whose identity changes over 1.0 second or 2.5 second periods, and concatenated speech of two speakers at a fixed point. The threshold α and F_0 changing ratio was set by referring the experiment results in 3.2, 3.3 and 3.4, and F_0 and spectrum were simultaneously modified. The speech, 24 in total, were presented to the 11 listeners at random, and they were asked to rate speech identity in 5 categories; i.e., "no difference in identity" gets 5 points and "large difference in identity" gets 1 point.

Figure 10 shows experiment results. If the acoustic difference between speakers is small, the 1.0 sec morphing period yields smooth changes. If the acoustic difference between speakers is large, it is necessary to lengthen the morphing interval.

The following audio files contain examples used in the experiment. [SOUND A261S01.WAV], [SOUND A261S02.WAV],[SOUND A261S03.WAV] contains speech for the left, middle and right speaker pair in Fig. 10, respectively. Morphing periods are from 1.45 sec to 3.85 sec, from 0.62 sec to 2.62 sec and from 0.84 sec to 3.73 sec, respectively.

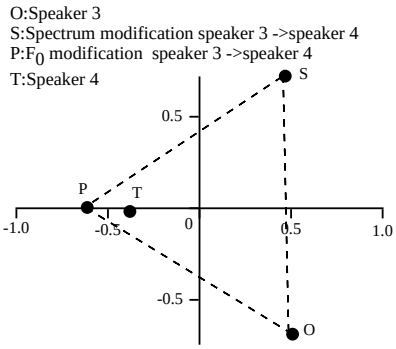


Fig. 8 Psychlogical distance of stimuli (3-4 pair)

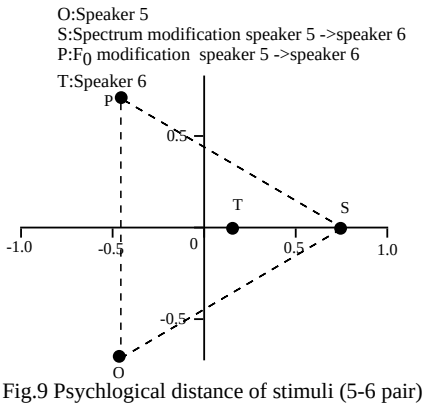


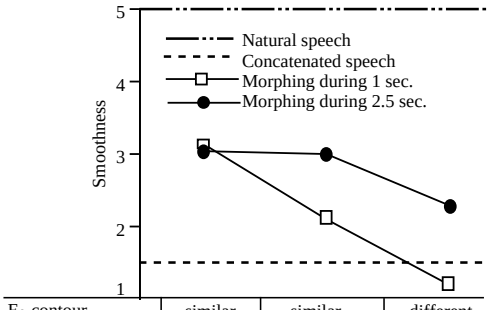
Fig.9 Psychlogical distance of stimuli (5-6 pair)

4. CONCLUSION

This paper proposed "speech morphing" as a new application of speech modification. The proposed algorithm has simple procedures such as spectrum mixing or overlap addition and experimental results show that the algorithm makes it possible to realize smooth changes in speech identity. Although, in this paper, we concentrated on smooth identity changes, speech morphing seems to have a much wider range of applications. These will be the future targets.

ACKNOWLEDGMENT

We are grateful to the members of the Speech Processing Department for their helpful discussions. We also thank Dr. Kitawaki, the department head, for his continuous support of this work.



F_0 contour	similar	similar	different
Average F_0 differece	middle(20H)	middle(20Hz)	large(50Hz)
Cepstrum distance	close(2.6dB)	far(4.3dB)	far(5.1dB)

Fig. 10 Smoothness of speech morphing

REFERENCES

[1] C. Hamon, E. Moulines, F. Chaarpentier, "A diphone synthesis system based on time-domain prosodic modification of speech", ICASSP'89 pp.238-241, 1989.

[2] C. Hayashi, "On the quantification of qualitative data from the mathematicostatistical point of view," Ann. Inst. Statist. Math 2, 1950.