A TIME-DOMAIN FEMALE-MALE VOICE CONVERSION ALGORITHM

LIU Li  YU Tiecheng
Speech Processing Laboratory, Institute of Acoustics, Chinese Academy of Sciences, Beijing
ll@speech1.ioa.ac.cn

ABSTRACT
In this paper, we put forward a time-domain female-male voice conversion algorithm. This method mainly focuses on two acoustic features that are thought to be the most important to speech individuality: pitch frequency and formant frequencies. To change pitch frequency, we cut off or add the low amplitude parts of speech signals in one pitch period. To change formants, according to the relationship between zero-cross rate and formants, and basing on the semi-waveform vector database which the former students formed during carrying out a speech waveform encoding algorithm, we use DTW technology to find a semi-waveform vector in the database to substitute the original semi-waveform. Experiments show that this algorithm is feasible. The average pitch frequency ratio of female speech to male speech is about 1.5 and the average formant frequencies ratio of female to male is about 1.2. We also found that the converted male voice is better than the converted female voice.

Key words: voice conversion, pitch frequency, formant frequency.

1. INTRODUCTION
Voice Conversion technology is a new technology that is coming up recently. Voice conversion is defined as modifying the speech signal of one speaker (source speaker) so that it sounds as if a different speaker (target speaker) had pronounced it. We can use this technology in many fields. In the speech synthesis fields, the output speech can be enriched using voice conversion. In addition, it can also be used in telephone voice translation, low-bit speech encoding, speaker adaptation, and so on. Voice conversion depends much on the research about voice individuality. Research indicate that the factors relevant to voice individuality can be distributed into two types. One is acoustic parameter, such as pitch frequency, formant frequencies and bandwidths, which are reflected by the voice source and the vocal tract of different peoples. The other is prosodic parameter, such as the timing, rhythm, and pause of voice, which usually depend on the social conditions of different peoples.

There have been many articles focusing on this topic [1]. Most of them are based on changing the acoustic parameters in frequency-domain. In this paper, we bring forward a new algorithm, which deal with speech waveform directly, to get the goal of changing voice individuality.

In section 2 we will discuss the modification of pitch and formant frequencies separately. In section 3, we present the system diagram and the experiment results. The summary is given in section 4.

2. TIME-DOMAIN PITCH AND FORMANTS MODIFICATION

2.1 Pitch Modification

2.1.1 Basic idea and algorithm
It is known that if we keep the zero-cross rate of speech signals changeless, then we can hold the main information of formant frequencies. And the speech signals with low amplitude make less effect on human's hearing, so we can cut off or add the last parts of speech signals during one pitch period, for they have the lowest amplitudes. Then the pitch period is shorten or lengthen accordingly, with keeping the formant frequencies changeless.

The algorithm is very simple. And the experiments show that our idea is right. The converted voice sounds that the pitch has already changed, and the waveform figure also proved it [3]. Though the pitch is changed, the voice quality of the new speech is changeless. We can hear clearly that the converted voice is the same word as the original voice. We also found
that though the pitch of the speech can be changed much, it sounds as if it had been articulated by the same gender. This indicates that the voice difference between male and female is not only reflected by the difference of pitch frequency, but also by that of formant frequencies.

From the spectrum of the original and new speech [3], we can see that the formants of converted speech are unchanged on the whole, so the voice quality is not changed. But the spectrum shows that the high-frequency parts of converted speech are larger than those of the source speech. It is because the algorithm has brought some noises to the speech. Though this made little effect to people’s hearing, we must improve our algorithm in the future.

2.1.2 Further discussion

Let us look at the process of human’s articulation to learn more about the speech waveform. After the vocal cords once vibrate, the glottal waves undulate and decrease in the vocal tract. The amplitude of the waveform envelope is reduced little by little. When the next stimulation happens, it goes the same progress again. The vocal cords of female are short and narrow, so their vibration is quicker than male. Thus the next stimulation happened quickly to make the waveform of new period. This phenomenon proves it is reasonable that we add or cut the low amplitude parts of speech signals in one pitch period to change the pitch. Fig.1 shows the phenomenon.

The former speech researchers found that the zero-cross rate of speech signals involved the information of formants. If the zero-cross position keep not changed, the intelligibility of speech signals can be hold with their amplitudes varying much. So we have the idea that the speech signals can be clustered into speech vector database according to the different length of semi-waveforms, then people’s voice can be reformed by the semi-waveforms in the vector database. Basing on this idea, the researchers in our lab have built a waveform encoding algorithm basing on semi-waveform vector quantification. This algorithm classified the speech waveforms into different types according to their length, and make vector quantification in turn. Experiments show that this algorithm have got good results. The MOS of the synthesized speech is above 3.5, so the algorithm can be used in speech transmission [2]. Fig.2 present the structure of the speech codebook.

2.2 Formants Modification

2.2.1 Semi-waveform vector database

Basing on the idea that changing the semi-waveforms of speech signals can change the characters of frequency-domain. We divided voiced sound parts of speech into a series of semi-waveforms, substituted each of them by the vector unit with different length of the speech vector database. To a semi-waveform with certain length, there have many samples in the database. In order to select appropriate one, We use DTW technology to find the most similar semi-waveform in the database to substitute original one.

From the spectrum of original and new speech signals [3], we can see that the formants of converted speech have already changed, and the ratio of semi-waveform length change is reciprocal of the ratio of the formant frequency change. This indicate the semi-waveform length of speech signal does reflect the spectrum characters.

![Fig.1 The waveform of the “ba4” articulated by male and female](image1)

![Fig.2 The structure of speech codebook](image2)
The hearing experiments show that the voice quality keep changeless on the whole when the change of formant frequencies is in a certain range. But out of the range, the voice quality changed much. The converted voice is not the same word as the original voice.

3. SYSTEM DIAGRAM AND EXPERIMENT RESULT

Fig.3 shows the framework of male-female voice conversion procedure.

![Diagram]

Fig.3 The framework of male-female voice conversion

Modulate the ratio $P$ and $F$ to make the converted voice sound best. For the speech conversion of female-to-male, $F$ is about 0.8, while $P$ is at 0.5. For the speech conversion of male-to-female, $F$ is about 1.2, while $P$ is at 1.5. Since we use the voice data of a single male and female, $P$ and $F$ may have some changes for other male and female, but will not change a lot. Thus we can conclude that the average pitch frequency ratio of female voice to male voice is about 1.5 and the average formant frequencies ratio of female voice to male voice is about 1.2.

The results of experiment show that the conversion of female-to-male is better than that of male-to-female. And the converted voice is not as good as what people actually speak. The converted female voice sounds a little sharp and the converted male voice sounds a little deep. It can be explained by the thoughts that besides pitch and formant frequencies, the parameters such as the glottal wave shape, spectral tilt shape, also have some influence on the variety of male and female voice.

4. SUMMARY

In this paper, we talked about a voice conversion algorithm. Since the pitch frequency and formant frequencies are thought to be the most important to speech individuality, the algorithm mainly focuses on these two parameters. To change pitch frequency, We cut off or add the low amplitude speech signals of a pitch period. And to change formants, according to the relationship between zero-cross rate and formants, and basing on the semi-waveform vector database, we use DTW technology to find a semi-waveform in the database to substitute the original semi-waveform. The results of experiment show that the average pitch frequency ratio of female speech to male speech is about 1.5 and the average formant frequencies ratio of female to male is about 1.2. And the converted male voice is better than the converted female voice. The advantage of this method is that the conversion of speech individuality is totally based on time domain, the algorithm is easy going, and the physical concept is clear and direct. This technology can be used in speech synthesis system and make the output speech more abundant.

5. REFERENCES

