VOICE CONVERSION THROUGH VECTOR QUANTIZATION

Masanobu ABE, Satoshi NAKAMURA, Kiyohiro SHIKANO, Hisao KUWABARA

ATR Interpreting Telephony Research Laboratories
2-1-61 Shiromi, Higashi-ku, Osaka, JAPAN

ABSTRACT

In this paper, we propose a new voice conversion technique through vector quantization and spectrum mapping. The basic idea of this technique is to make mapping codebooks which represent the correspondence between different speakers' codebooks. The mapping codebooks for spectrum parameters, power values, and pitch frequencies are separately generated using training utterances. This technique makes it possible to precisely control voice individuality.

To evaluate the performance of this technique, hearing tests are carried out on two kinds of voice conversions. One is a conversion between male and female speakers, the other is a conversion between male speakers. In the male-to-female conversion experiment, all converted utterances are judged as female, and in the male-to-male conversion, 65% of them are identified as the target speaker.

1. INTRODUCTION

In daily communication, voice individuality is one of the most important aspects of human speech. It is especially important to identify the speaker when a conversation is made through a telephone line. A technique to control speech individuality, therefore, plays an important role and offers a lot of applications. Our present study is concerned with converting voice quality from one speaker to another and developing a technique which enables us to give individuality to synthesized speech.

Speech individuality generally consists of two major factors. One is acoustic features and the other is prosodic features. As the first step in this research, we are trying to control the acoustic features[1][2]. According to previous studies, the acoustic features that contribute to speech individuality are distributed among various parameters, such as formant frequencies, formant bandwidths, spectral tilt, and glottal waveforms[3][4].

In this paper, we propose a new voice conversion technique without separating these acoustic parameters. The basic idea of this technique is to make use of codebooks for several acoustic parameters. These codebooks carry all information about the speech individuality in terms of the varying acoustic features. A conversion of acoustic features from one speaker to another is, therefore, reduced to the problem of mapping the codebooks of the two speakers. In section 2, a method of making mapping codebooks and a procedure of synthesis are described. In section 3, the mapping codebooks are evaluated by measuring distortion. In section 4, the performance of this technique is evaluated as a whole by hearing tests.

2. VOICE CONVERSION THROUGH VECTOR QUANTIZATION

Our voice conversion technique consists of two steps: a learning step and a conversion-synthesis step. The learning step is a process to generate the mapping codebooks, and the conversion-synthesis step is a process to synthesize speech using the mapping codebooks.

2.1 Learning step

The mapping codebooks are codebooks that describe a mapping function between the vector spaces of two speakers. The block diagram in Fig.1. illustrates how a mapping codebook for spectrum parameters is generated.

1.Two speakers, A and B, pronounce a learning word set. Then all words are vector-quantized frame by frame.

2.The correspondence between vectors of the same words from the two speakers is determined using Dynamic Time Warping(DTW). This is done for all the learning word set.

3.The vector correspondences between two speakers are accumulated as histograms.

4.Using each histogram as a weighting function, the mapping codebook is defined as a linear combination of speaker B's vectors.

5.Steps 2, 3, and 4 are repeated to refine the mapping codebook.

Mapping codebooks for pitch frequencies and power values are also generated because these parameters contribute a great deal to speech individuality. These mapping codebooks are generated at the same time

CH2561-9/88/0000-0655 $1.00 © 1988 IEEE

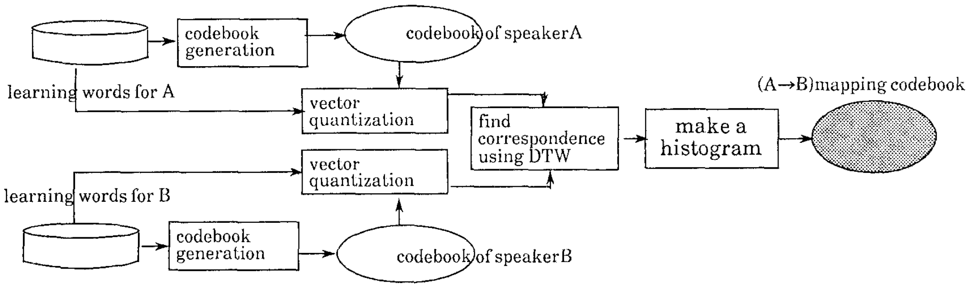


Fig.1. Method for Generating a Mapping Codebook

according to almost the same procedure mentioned above. The differences are

1.Pitch frequencies and power values are each scalar-quantized.

2.The mapping codebook for pitch frequencies is defined based on the maximum occurrence in the histogram.

2.2 Conversion-synthesis step

As shown in Fig.2. after speaker A's speech is analyzed by the linear prediction method, the spectrum parameters are vector-quantized using his/her codebook, and parameters for pitch frequencies and power values are scalar-quantized using his/her codebooks. Then, synthesis is carried out by decoding them using mapping codebooks between speakers A and B. The output speech will have the voice individuality of speaker B.

3. CONVERSION EXPERIMENTS

To evaluate the performance of the conversion technique, distortion measurements were carried out the spectrum parameters as well as the pitch frequencies.

3.1 Spectrum conversion experiments

Experimental conditions are listed in Table 1. A set of 100 phonetically-balanced words was used in this experiment. Spectrum conversions were made between female and male voices, between male and male and between female and female voices. Voices of six speakers (3 male and 3 female speakers, all professional announcers) were used as speech material.

Table 2 shows the results of the open test. After vector-quantization, two kinds of spectrum distortions between two speech samples were calculated:between the input and the target speaker's (before conversion in Table 2), and between the converted speech and the target speaker's speech. In the female-to-female conversion, the distortion decreased by 27% compared to non conversion, by 49% for the male-to-male conversion, and by 66% for the male-to-female conversion.

3.2 Pitch frequency conversion experiments

Pitch frequency conversion was also carried out through the same process described in 3.1. The experimental results are shown in Fig.3. This figure shows the relation between the number of learning words and the average pitch frequency differences after conversion. The value at the point where the number of learning words is 0 shows the natural average pitch frequency difference between the two speakers.

According to this figure, 60 words are considered to be enough to make a mapping codebook for pitch frequency regardless of speaker combinations, and the average pitch frequency differences decrease down to less than 15Hz.

4. EVALUATION BY HEARING TEST

To evaluate overall performance of this technique, three kinds of hearing tests were carried out. The first experiment deals with the male-to-female conversion and the other two experiments deal with the male-to-male conversion.

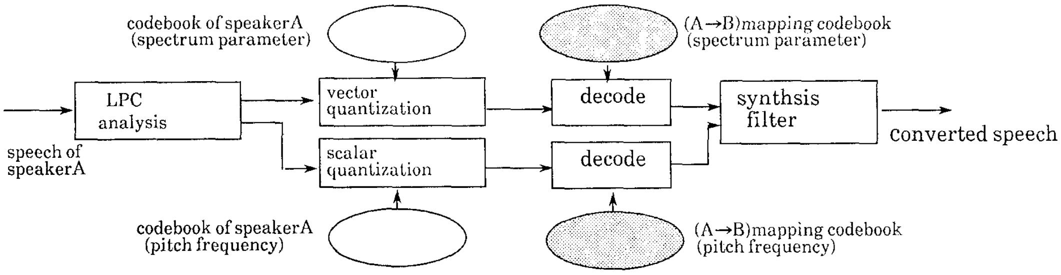

codebook of speakerA (spectrum parameter)
(A→B)mapping codebook (spectrum parameter)
LPC analysis
vector quantization
decode
synthesis filter
converted speech
speech of speakerA
scalar quantization
decode
codebook of speakerA (pitch frequency)
(A→B)mapping codebook (pitch frequency)

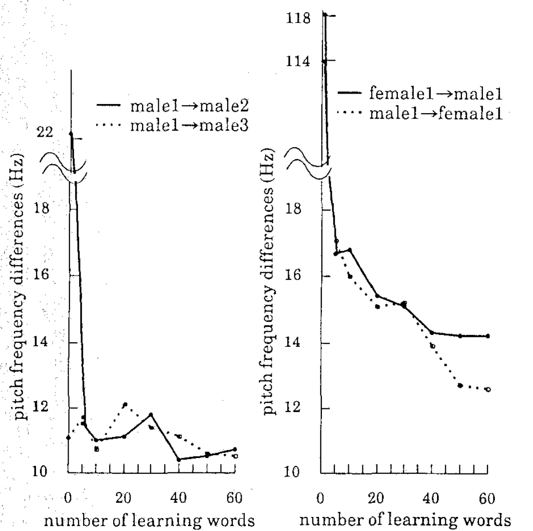
Fig.2. Block Diagram of a Conversion from Speaker A to Speaker B

Table 1 Experimental Conditions

A/Ddata	12KHz sampling、16bit
window length	256points (21.3msec)
window shift	36points (3.0msec)
analysis order	12
clustering measure	WLR
learning samples for clustering	5000 frames
codebook size for spectrum parameter	256
learning words for mapping	100 words
codebook size for pitch frequency	35 ~ 64

Table 2 Spectrum Distortion

speaker combination	before conversion	after conversion
female1→female2	0.2759	0.2109
female1→female3	0.2070	0.1489
male1→male2	0.3364	0.1717
male1→male3	0.2851	0.1550
male1→female1	0.6084	0.2193



1. male→ male conversion 2. male←→fmale conversion

Fig.3. Pitch Frequency Differences for Number of Learning Words

4.1 Experimental Procedure

4.1.1 Experiment 1

Experiment 1 was designed to evaluate the voice quality for male-to-female voice conversion by a pair-comparison hearing test. In addition to the fully converted speech, conversion was also done for pitch and spectrum parameters separately in order to examine the individual contribution of these parameters to speech individuality. The following is a list of 5 different speech conversions performed in this experiment.

1.vector-quantized original male speech(m)

2.male-to-female converted speech; pitch frequency conversion only(mp-fp)

3.male-to-female converted speech; spectrum conversion only(ms-fs)

4.male-to-female converted speech on all parameters(m-f)

5.vector-quantized original female speech which is the target for the conversions(f)

In order to avoid unnecessary cues for the judgment of voice quality, 2 different words were used to make speech pairs for the hearing test. A set of speech pairs consists of all possible combinations of stimuli from the 5 different conversions, 40 in total. They were presented to listeners through a loud-speaker in a sound-proof room. Twelve listeners participated in the experiment and they were asked to rate the similarity for each pair on five categories: "similar","slightly similar","difficult to decide","slightly dissimilar","dissimilar".

4.1.2 Experiment 2

Experiment 2 was designed to evaluate the conversion between two male speakers by the ABX method. Stimuli A and B are vector-quantized original speech tokens for speakers N and M respectively. X is either the N→M or M→N converted token. Four different words were used for the conversions and each triad was a combination of 3 different words. As a total, 48 speech triads were presented to the listeners in the same way as described above. The listeners were required to select the stimulus (A or B) which most closely resembled stimulus X in speaker identity.

4.1.3 Experiment 3

Experiment 3 was designed to evaluate the conversion between male speakers in the same way as in 4.1.1. However, conversions for pitch frequencies alone and spectrum parameters alone were excluded. The following is a list of the 4 conversions.

1.vector-quantized male speech (male1)

2.same as 1 but for another male speaker (male2)

3.converted speech from male1 to male2 (male1→male2)

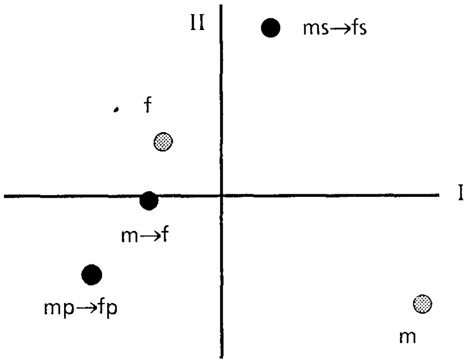
4.converted speech from male2 to male1 (male2→male1)

In total, 72 speech pairs were generated and the experimental procedures were the same as in experiment 4.1.

4.2 Experimental results

4.2.1 Evaluation on male-to-female conversion (Results of Experiment 1)

Hayashi's forth method of quantification[5] was applied to the experimental data obtained by the hearing test. This method is to place stimuli on a space according to the similarities between every two stimuli,

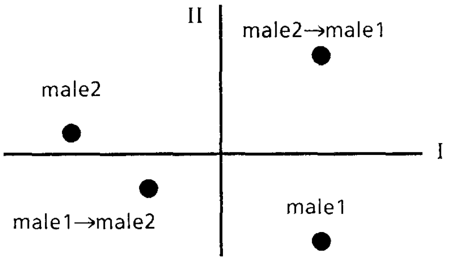


m: vector-quantized original male speech
f: vector-quantized original female speech
mp→fp: male-to-female converted speech only for pitch frequency
ms→fs: male-to-female converted speech only for spectrum parameters
m→f: male-to-female converted speech for all parameters

Fig.4. Distribution of Psychological Distances for the Male-to-Female Conversion

Table 3 Percentages of Correct Responses

speaker combination	correct answer response(%)
male 1→male 2	64.6
male 2→male 1	63.6
male 1→male 3	58.0
male 3→male 1	56.8



male1: vector-quantized male 1 speech
male2: vector-quantized male 2 speech
male1→male2: converted speech from male 1 to male 2
male2→male1: converted speech from male 2 to male 1

Fig.5. Distribution of Psychological Distances for the Male-to-Male Conversion

and its formulation is to minimize the measure Q, where

$$Q = -\mathrm{SUM}[e(i,j)\{x(i)-x(j)\}^2]$$

e(i,j) denotes the similarity between stimuli i and j, and x(i) represents the location of stimulus i in the space. The projection onto a two-dimensional space is shown in Fig.4. It represents the relative similarity-distance between stimuli. In this figure, converted speech "m→f" is placed most closely to the speech "f". This indicates that the male speech was properly converted to the target female speech by this technique. Judging from the positions of "mp→fp" and "ms→fs", it is observed that the first and second axes roughly correspond to pitch frequency and spectrum differences, respectively. The result indicates that neither pitch frequency nor spectrum carries enough information about speech individuality, and both are necessary.

4.2.2 Evaluation on male-to-male conversion (Results of Experiment 2 and 3)

The result from Experiment 2 is shown in Table 3. The numbers in this table represent the percentage of response where converted speech samples are judged as the target speaker's. The difference in percentage between two speaker-combinations is caused by average pitch frequency differences between those speakers as shown in Fig.3. Fig.5. represents the result for Experiment3 analyzed by the same method as in 4.2.1. It is observed that the converted speech samples, "male 1→male 2" and "male 2→male 1", are both placed closer to their target speech.

5. CONCLUSION

We proposed a new voice conversion technique through vector quantization and spectrum mapping. The advantage of this technique is summerized as follows;

1.The mapping codebooks which make it possible to give an individuality to synthesized speech are generated for each speaker from training on a limited number of word utterances.

2.The mapping codebooks enable voice conversion between any two speakers with high quality.

3.In the synthesis process, only a small amount of computation is required.

From the hearing test, we can conclude that the converted speech has a voice quality very close to the target speaker's voice.

Acknowledgments

We are grateful to Dr. Tohkura for helpful discussion. We wish to also thank Dr. Kuremastu for his continuous support of this work.

References

[1]Shikano,K. , Lee,K. , Reddy,R. , "Speaker Adaptation Through Vector Quantization", ICASSP,pp. 2643-2646, April 1986.

[2]Nakamura,S. , Shikano,K. ,"Spectrogram Normalization Based on Vector Quantization"(in Japanese), Trans. of the Committee on Speech Research, The Acoustical Society of Japan, Vol. SP87-17, pp. 9-16, June 1987, .

[3]Kuwabara,H. , Takagi,T. ,"Quality Control of Speech by Modifying Formant Frequencies and Bandwidth", 11th Inter. Congress of Phonetic Sciences, pp. 281-284, August 1987.

[4]Childers,D.G. , Yegnanarayana,B. , Wu,K. ,"Voice Conversion: Factors Responsible for Quality", ICASSP, pp. 748-751, March 1985.

[5]Hayashi,C. ,"Recent Theoretical and Methodological Developments in Multidimensional Scaling and its Related Methods in Japan", Behaviormetrika, No.18.