TR-I-0126

言語にわたる声質変換
Cross-Language Voice Conversion

阿部 匡伸
Masanobu ABE

(ATR自動翻訳電話研究所)
ATR Interpreting Telephony Labs.

1989.12

内容梗概

言語わたる声質変換の最終目的は、自動翻訳された音声(英語)に話者(日本人)の個人性を付与することである。基本検討として、日本語と英語のバイリンガル話者の音声を分析して、以下のことが明らかとなった。

- バイリンガルの発声した英語と日本語は、2名の日本人の発声した日本語と比べて、スペクトルの広がりは小さく、スペクトルの分布も近い。
- 英語に特徴的なスペクトルは、/r/、/æ/、/č/、/š/ などの限られた音韻に現れる。
- 英語の静的なスペクトル特徴は上述のように明らかであるが、聞こえの点ではその差は少ない。

この結果をもとに言語にわたる声質変換モデルを考を提案し、変換実験を行った。その結果、英語の音韻了解性を保存したまま日本人の声質を得ることができた。また、言語にわたる声質変換モデルのための尺度を提案し、有効に使用できることを確認した。

Cross-Language Voice Conversion

Masanobu ABE

November 22, 1989

Abstract

The goal of cross-language voice conversion is to preserve the speech characteristics of one speaker when that speaker's speech is translated and used to synthesize speech in another language. Concerning this goal, we investigate two issues in this paper. The first is confirming the spectrum difference between English and Japanese. The experimental results using a bilingual speaker's speech data are the following: (1)Inter-language(between English and Japanese) difference is smaller than inter-speaker differences, and (2)Judging from listening tests, the difference between English and Japanese is very small. The second issue is a model for a cross-language voice conversion. In our approach, voice conversion is considered a mapping problem between two speakers' spectrum spaces. From this point of view, we propose a voice conversion model and measures for the model. The converted speech from male to female is as understandable as the unconverted speech, and moreover it is recognized as a female's voice.

1 Introduction

In recent years international communication has been increasing all over the world, and we have a lot of opportunities to communicate with foreigners using other languages. Under this situation, we have to make an effort to communicate in other languages, sometimes making mistakes in understanding or, if we don't know the language, not understanding anything at all. A system developed to overcome such a language barrier by making use of the latest information processing technology would be very useful.

One of these systems guided by the motivation above is an automatic telephone interpretation system: i.e, a facility which enables a person speaking in one language to communicate readily by telephone with someone speaking another language[1]. A block diagram of the system is shown in Fig.1. The system consists of three constituent technologies: speech recognition, machine translation and speech synthesis. When a person speaks Japanese at one end of a telephone line, for example, the speech recognition subsystem recognizes his/her speech, the translation subsystem translates Japanese spoken dialogue into English, the speech synthesis subsystem synthesizes English speech, and then a listener hears English at the other end of the line.

To develop an interpreting telephone, there are many issues to be solved in each subsystem. The theme which is investigated in this paper is speech individuality control. By individuality control, we mean the generation of inteligible speech while maintaining the personel characteristics of the original speaker. In daily communication we have much experience with speech individuality. When we converse by telephone, for example, speech individuality makes it possible for us to identify who is talking over the telephone. In the case of an interpreting telephone, as shown in Fig.1, the output speech is not uttered by the speaker but synthesized speech. That means it is necessary to give speech individuality to the synthesized speech, and moreover speech individuality should be given to speech uttered in a different language from the language that the speaker speaks. Therefore the ultimate goal of the speech individuality control is to convert speech quality from one speaker who speaks a language to another speaker who speaks another language. We call it cross-language voice conversion.

The cross-language voice conversion problem is separated into two subjects; one is how to control speech individuality, the other is how to solve the problems in cross language synthesis. In this paper the subjects are reported as follows. In section 2, in terms of speech individuality control, a voice conversion method based on vector quantization is described. In section 3, in terms of the cross language problem, the spectrum difference between Japanese and English is investigated. In section 4, integrating the result of section 2 and 3, a cross-language voice conversion model and measures for

the model are proposed.

2 Voice conversion through vector quantization

2.1 Basic idea

There are two aspects of speech spectrum characteristics; one is static (frame-wise) characteristics and the other is dynamic characteristics. At this stage, we are concerned with how to control the first one. According to previous studies, the static spectrum characteristics that contribute to speech individuality are formant frequencies, formant bandwidths, spectral tilt, and glottal waveforms[2,3]. Because speech individuality is determined by all of these parameters, it is difficult to control voice quality by modifying each parameter independently.

On the other hand, codebooks used in vector quantization represent spectrum characteristics that all of these parameters contain. Therefore, it is possible for code vectors in codebooks to represent speech individuality. A conversion of the static spectrum characteristic of one speaker to that of another is reduced to the mapping problem for the two speakers' codebooks. We have already proposed a voice conversion method based on the idea, and confirmed good performance of voice conversion between Japanese speakers[4]. The algorithm is explained briefly in the following section.

2.2 The algorithm to generate mapping codebook

In the voice conversion method, a mapping function between the vector space of two speakers is represented by a "mapping codebook." The block diagram in Fig.2 illustrates how a mapping codebook is generated using training data. The training is performed as follows:

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 training 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. Step 2,3 and 4 are repeated to refine the mapping codebook.

The codebooks are generated by the LBG algorithm[5]. Analysis conditions are shown in Table 1.

3 Japanese spectrum space vs. English spectrum space

Let's suppose that there is a person who can only speak Japanese and has no knowledge of English, and we have a lot of speech uttered by him. Then using the speech signal, is it possible to produce English which sounds like his speech? If English speech were constructed from the same spectral pool as Japanese, it should be possible to find the proper spectra and rearrange them for English. If not, it is necessary to generate or estimate certain spectra of English which do not exist in Japanese. If so, which spectra are they? When the voice conversion method explained in section 2 is applied to cross-language voice conversion, it is necessary to address these questions. This is the topic of this section.

3.1 How large is the spectrum difference between Japanese and English?

3.1.1 Experimental method

To investigate the spectrum difference between Japanese and English we use vector quantization, because the experimental results are easily applied to voice conversion based on vector quantization. As analysis data, we collected

speech uttered by a bilingual speaker. Using these data makes it possible to eliminate the speaker individuality factor, and study only language differences. The bilingual speaker was born in Japan, lived in Switzerland from age 2 to 4, and after that was brought up in Japan. His mother is Japanese and his father is German. He went to international school in Tokyo where all lessons were held in English. According to native speaker' judgments, his English and Japanese pronunciation is as good as native speakers'. He uttered 216 Japanese words and 328 English words. The English words are shown in Appendix A. Both word sets were selected to be phonetically balanced.

A way to know how large the relative spectrum difference is between different languages is to compare the spectrum difference between speakers of the same language. Moreover it is easy to understand, because there are many studies on inter-speaker difference. Therefore, we also collected speech data uttered by three Japanese male speakers and three Japanese female speakers. Then seven codebooks were generated by the LBG algorithm for the following data sets.

(1)English and Japanese words uttered by the bilingual speaker
(2)English words uttered by the bilingual speaker
(3)Japanese words uttered by the bilingual speaker
(4)Japanese words uttered by one male and one female speaker
(5)Japanese words uttered by two male speakers
(6)Japanese words uttered by two female speakers
(7)Japanese words uttered by one male speaker

After vector quantizing each utterance, an occurrence number of each codeword in the following category pairs was counted up;

(A)English vs. Japanese
(B)male speaker vs. female speaker
(C)male speaker1 vs. male speaker2
(D)word set1 vs. word set2 (set1 and set2 uttered by the same speaker)

To measure the distribution distance between the above categories, Kullback's divergence was calculated. Kullback's divergence is defined as follows;

$$D = \sum_{i=1}^{r} [P(a_i|\omega_1) - P(a_i|\omega_2)] \log \frac{P(a_i|\omega_1)}{P(a_i|\omega_2)}$$

Here,
ω_1:category 1
ω_2:category 2
$P(a_i|\omega_j)$:posteriori probability of the codeword a_i in category ω_j.

3.1.2 Experimental results and discussion

Fig.3 shows spectrum distortion of (1),(2),(3) and (7) in 3.1.1 according to codebook size. Because the spectrum distortion in (2) and (3) is almost the same, it says that the size of Japanese and English spectra is almost the same. On the other hand, the distortion of (1) in an 8-bit codebook is as much as the distortion of the others in 7-bit codebooks. That means that the size of the spectrum space is doubled when Japanese and English are mixed.

Fig.4 shows spectrum distortion of (1),(4),(5) and (7) in 3.1.1 according to codebook size. The distortion in (1) is smaller than the distortion in (4) and (5). Therefore, when Japanese and English are mixed, the size of the spectrum space is not so large as the spectrum space size of two speakers.

Table 2 shows Kullback's divergence for each category pair. Kullback's divergence indicates the overall distance between two distributions. Therefore, the larger the value is, the better two categories are separated. Table 2 shows that the data uttered by the male speaker and the female speaker are well separated, and that data uttered by the same speaker can hardly be separated. Judging from the value of the English-Japanese pair, the two categories show overlap more than separation. To show the fact visually, scatter plots are shown in Figs. 5,6,7,8. Points in the figures show codeword vectors, and they are plotted according to the occurrence number in each category.

To summarize the results, in terms of the Japanese and English spectrum space, the size is smaller than a two-speaker spectrum space, and its dis-

tribution is closer than that of a two-speaker spectrum space. Fig.9 shows an outline of the summary. The distance between peaks in the upper figure shows the distribution distance; total area of the lower figure shows the distortion.

3.2 Which spectra are unique to English?

The results in 3.1 indicate that some spectra only exist in English. In this section, it is shown which phonemes contain such spectra.

3.2.1 Experimental method

The alignment of a phonetic transcription for English uttered by the bilingual speaker is performed by CASPAR, an automatic alignment system developed at MIT[6]. Some errors are corrected by hand. After that, occurrence numbers of each phonetic segment are counted up for each codeword of the codebook generated from utterrances spoken by the bilingual English and Japanese speaker.

3.2.2 Experimental results and discussion

The spectrum characteristics of codewords which frequently occurred in English but not so often in Japanese are summarized as follows. Figures referred to in each description show an LPC spectrum envelope and a histogram. The histogram shows the occurrence percentage of a transcribed segment in a codeword. As a reference, the F1-F2 relationship for vowels in English and Japanese is shown in Fig.10[7,8].

1. Vowels /ɔ/, /ə/, /ɑ/. F1 and F2 is very close. This formant structure is never found in codewords which frequently occurred in Japanese. (Fig. 11)

2. Vowel /æ/. This is the typical formant structure of /æ/. As shown in Fig.10. Japanese has no vowel of this formant structure. (Fig.12)

3. Consonants /š/, /ǰ/,/č/. These are voiceless consonant but formant structure is very clear in spectrum envelope. Voiceless consonant code-

words which frequently occurred in Japanese do not have such a clear formant structure. (Fig.13)

4. Consonants /f/, /t/. These are voiceless consonants but formant structure is very clear.(Fig.14)

5. Liquid /r/. F2 and F3 are very close. This is the typical /r/ in English.(Fig.15)

6. Vowel /ɪ/. This is the typical formant structure of /ɪ/. Japanese has no vowel with this formant structure. (Fig.16)

3.3 How does the difference sound?

The results in 3.1 and 3.2 show that there is a spectral difference between Japanese and English and the difference is observed in particular phonemes. In this section we examine the differences between English and Japanese using a perceptional experiment.

3.3.1 Experimental method

The English speech uttered by the bilingual speaker is synthesized in the following two ways. One synthesized speech is coded by English, then decoded by English(CEDE), and the other is coded by Japanese, then decoded by Japanese(CJDJ). Twenty eight word sets which contain all phonemes more than once are synthesized using the extracted pitch frequency and speech power. The CEDE and CJDJ pair of the synthesized words are presented to 8 listeners(4 male, 4 female) using a headphone and they are asked to judge as follows:

Is there a difference between the pairs? (1)If no, go to next word. (2)If yes, indicate the better one and give a reason for the choice.

3.3.2 Experimental results and disccusion

Table 3 shows the percentage of times that the distinction between the two words was judged correctly, incorrectly or the two words were judged indistinguishably. Here, we use the term "correct" when CEDE is judged better

than CJDJ. The answers mostly depend on words. Some words have a tendency to be judged indistinguishable, some words are judged correctly half the time, some words are usually judged correctly. In Table 4, the words used in the experiment are classified into the three categories. The fact that half of the words are judged indistinguishable is reasonable according to the results in 3.1. The phonemes that are judged better sounding are /æ/,/ə/,/č/,/š/,/ŋ/,/r/. The results agree with the results in 3.2. However none of the words are judged perfectly as the same category. That means that the difference between CEDE and CJDJ is very small.

4 Cross-language voice conversion

4.1 Cross-language voice conversion model

The voice conversion method explained in section 2 needs training speech data uttered by two speakers to generate the mapping codebook. As mentioned in section 3.3, the speech coded by a Japanese codebook sounds almost the same as the speech coded by an English codebook. Therefore, in terms of the training speech data, the first hypothesis we have in the cross-language voice conversion is that the spectrum correspondence between Japanese and English should be found if both speech sounds are the same.

Moreover, we have the other hypothesis that synthesized speech by English synthesis-by-rule systems preserves spectrum characteristics of English even if the input string or duration is modified a little, because in the case of interpreting telephone, as mentioned in the introduction, we would like to preserve a speaker's individual characteristics in the synthesized speech. Following from these two hypotheses, Japanese words are synthesized using MITalk system[9] as the training data. The block diagram of the cross-language voice conversion system is shown in Fig.17.

4.2 Preliminary experiment

Two kinds of synthesis-by-rule systems(MITalk-E, MITalk-Ed) were used. MITalk-E synthesizes speech using a spelling which makes synthesized speech sound like Japanese. In addition to MITalk-E, MITalk-Ed synthesizes speech

using phoneme durations extracted from Japanese speaker's utterances. One hundred words were uttered by six Japanese speakers(three male and three female) and synthesized by MITalk-E and MITalk-Ed. The mapping codebooks are generated for all speaker pairs by the algorithm explained in section 2.

The results were very impressive. The converted speech was as understandable as the MITalk speech. And also in the case of the conversion between male and female speech, the converted speech was recognized as female speech. But the performance of cross-language voice conversion was subjectively judged to be slightly worse than voice conversion from Japanese.

4.3 Measures for cross-language voice conversion model

The result in section 4.2 shows that cross-language voice conversion is more difficult than voice conversion within the same language. To improve the cross-language voice conversion, the distortion measure is not enough. In this section, some measurement criteria for cross-language voice conversion are investigated.

4.3.1 Mutual information

Voice conversion through vector quantization is considered as an information channel as shown in Fig.18. An input alphabet A = $\{a_i\}$, i=1,2,...,r are the codewords of speaker A, and output alphabet B = $\{b_j\}$, j=1,2,...,r are the codewords of speaker B. Because the speakers' spectrum spaces are different from each other, a codeword of speaker A doesn't always have a one-to-one correspondence with a particular codeword of speaker B. This uncertainty is measured by mutual information $I(A;B)$.

$$I(A;B) = H(A) - H(A|B)$$

Here,

$$H(A) = \sum_{i=1}^{r} P(a_i) \log \frac{1}{P(a_i)}$$

$$H(A|B) = \sum_{i=1}^{r} P(b_i) \sum_{j=1}^{r} P(a_i|b_j) \log \frac{1}{P(a_i|b_j)}$$

$P(a_i)$:probability of codeword a_i of speaker A
$P(a_i|b_j)$:posteriori probability of the input symbol a_i

4.3.2 Entropy of Speaker Markov model

Mutual information can only deal with static characteristics of speakers. To make use of the dynamic characteristics, a speaker Markov model is used. The speaker Markov model is shown in Fig.19[10]. The states of this model are the codewords of the target speaker. Transitions are possible from each state to every other state. The output is a codeword of the original speaker. Dynamic characteristics are measured by the entropy of the Speaker Markov model which is calculated as follows;

$$H(SM) = \sum_{j=1}^{r} P(b_j) H(A|b_j)$$

$$H(A|b_j) = \sum_{i=1}^{r} P(a_i|b_j) \log \frac{1}{P(a_i|b_j)}$$

$$P(a_i|b_j) = \sum_{k=1}^{r} P(t_{jk}) P(a_i|t_{jk})$$

$P(b_i)$:probability of codeword b_i of speaker B
$P(a_i|b_j)$:a posteriori probability of the output symbol a_i
$P(t_{jk})$:transition probability from b_j to b_k
$P(a_i|t_{jk})$:probability of a_i when transition is from b_j to b_k

4.3.3 Experimental results and discussion

Both measures are calculated for all data in 4.2. Fig.20 shows mutual information and Fig.21 shows entropy of speaker model for each speaker pair, i.e., Japanese vs. Japanese(J-J), Japanese vs. MITalk-E(J-E), Japanese vs.

MITalk-Ed(J-Ed), MITalk-E vs. MITalk-Ed(E-Ed). Moreover, in these figures, the values for the above four combinations are separately shown for the male-male pair and male-female pair.

As discussed in 3, the inter-language difference is smaller than inter-speaker differences in terms of the spectrum distortion. However the performance of the voice conversion in the inter-language case is not as good as in the inter-speaker case. Therefore, the most important property of measures investigated in this section is the ability to distinguish other differences between inter-speaker and inter-language. Fig.20 and Fig.21 show that the mutual information and entropy of the speaker model have larger differences in the inter-language case than the inter-speaker case. Therefore, both measures are useful to evaluate the differences between languages.

Any measurement also must preserve the inequality of the objects. In other words the closer two objects are, the less the value of the measurements should be. In the experiments, we can say that the J-J pair is closest, the E-J pair is farthest and the J-Ed and E-Ed pairs are midway between the two. The reasons are as follows; (1)One of the biggest difference between Japanese and English is phoneme duration. Because MITalk-Ed is given Japanese phoneme duration, the correspondence between the J-Ed pair is more consistent than that of the E-J pair. (2)Because MITalk-E and MITalk-Ed have the same rule of spectrum pattern generation, the distortion measure is more reliable in E-Ed pair than in E-J. From the inequality preservation point of view, Fig.20 and Fig.21 show that both the mutual information and entropy of the speaker model have an adequate property for measurement.

In comparing mutual information(MI) and entropy of speaker model(ESM), ESM can distinguish the difference between J-Ed and E-Ed, but MI can not. The reason ESM is superior to MI is caused by the quantity of information when they are calculated, i.e. ESM needs not only the correspondence of codewords at a particular time, but also the codewords at one unit time before.

5 Conclusions

We proposed the idea of cross-language voice conversion, and obtained the following results.

- The spectrum differences between English and Japanese are studied by comparing the spectrum difference among different speakers. The results show its size to be smaller than two-speaker spectrum space, and its distribution to be closer than that of a two-speaker spectrum space.
- The spectrum which characterized English appeared in particular phonemes. They are /ɔ/, /ə/, /ɑ/,/æ/,/š/, /ǰ/,/č/,/f/, /t/,/r/,/ɪ/.
- English words were synthesized without using the spectrum which characterized English. Judging from listening tests, the speech sounds are close to English.
- Cross-language voice conversion was performed using a voice conversion method based on vector quantization. The converted speech was as understandable as the MITalk speech. And also in the case of the conversion between male and female speech, the converted speech was recognized as female speech.
- Mutual Information and entropy of the Speaker Markov model are proposed as measurement criteria of cross-language voice conversion. Experimental results show the measures work well.

To summarize, the cross-language voice conversion model described in this paper works well as a first approximation. The key point to improve the performance is the method to consistently find the correspondence between the codewords. In this case, as investigated in section 4, Mutual information and entropy of the speaker model are very useful as measures. The other approach for cross-language voice conversion is to estimate the codewords which characterize English. In the future, we would like to investigate these points.

6 Acknowledgments

The work described in this paper was carried out while the author was staying at MIT Laboratory for Computer Science from June to November 1989 as a visiting scientist. The author is very grateful to Dr. Victor W. Zue for providing the opportunity to work for him and for his discussion on this work. The author also appreciates members of SLS for giving a comfortable environment to pursue the work and for cooperating on the listening test. Finally, the author appreciates Dr. Kurematsu and Dr. Shikano for encouragement and continuous support.

References

[1] Kurematsu, H, "Automatic Telephone Interpretation: A Basic Study",ATR Technical Report, May 1987.

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

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

[4] Abe, M, Nakamura S, Shikano K, Kuwabara H, "Voice Conversion Trough Vector quantization," ICASSP, pp.655-658, April 1988.

[5] Linde, Y, Buzo A, Gray R M, "An Algorithm for Vector Quantizer Design," IEEE Trans. on Communication,Vol.COM-28, pp.84-95, January 1980.

[6] Leung, C L, Zue W V, "Automatic Alignment of Phonetic Transcriptions with Continuous Speech," IASTED International Symposium on Robotics and Automation, pp.24-27, June 1985.

[7] Zue, W V, "Speech spectrogram reading," MIT special summer course lecture note, July 1985.

[8] Umeda, N, NTT Electrical Communication Labs. Technical Report,No.579, July 1957.

[9] Allen, J, Hunnicutt S, "MITalk-79:The 1979 MIT Text-to-speech system," Proc. of the 97th Meeting of ASA, YY1, June 1979.

[10] Rigoll, G, "Speaker adaptation for large vocabulary speech recognition systems using 'speaker Markov models'," ICASSP, pp.5-8, May 1989.

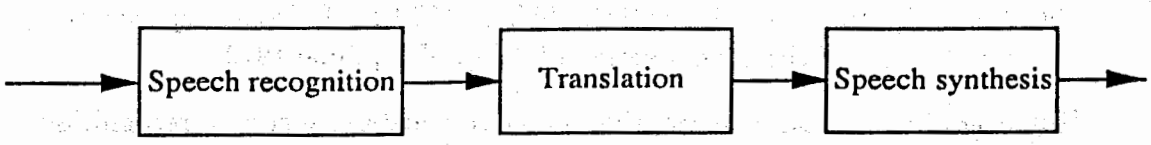


Fig.1 An interpreting telephone

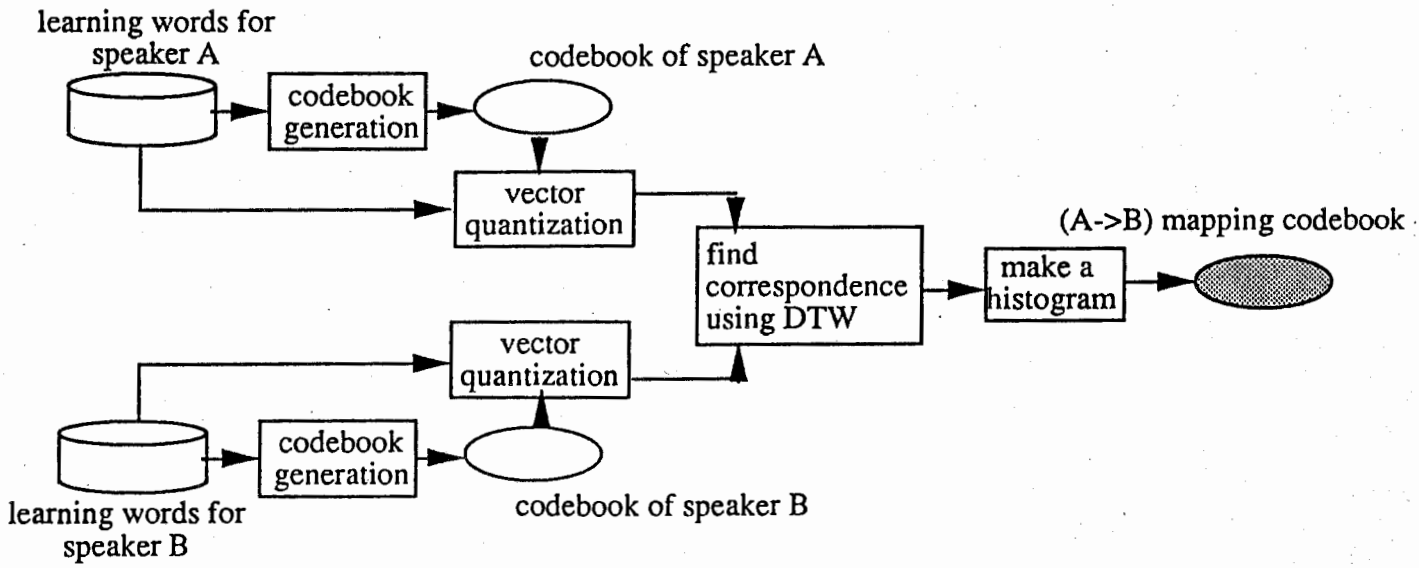


Fig.2 An algorithm for generating a mapping codebook

Table 1 Experiment conditions

A/D data	12kHz sampling, 16bit
window length	256points(21.3msec)
window shift	36points(3msec)
LPC analysis order	14
clustering measure	WLR
samples for clustering	12000frams
codebook size	256
training words number for mapping	100

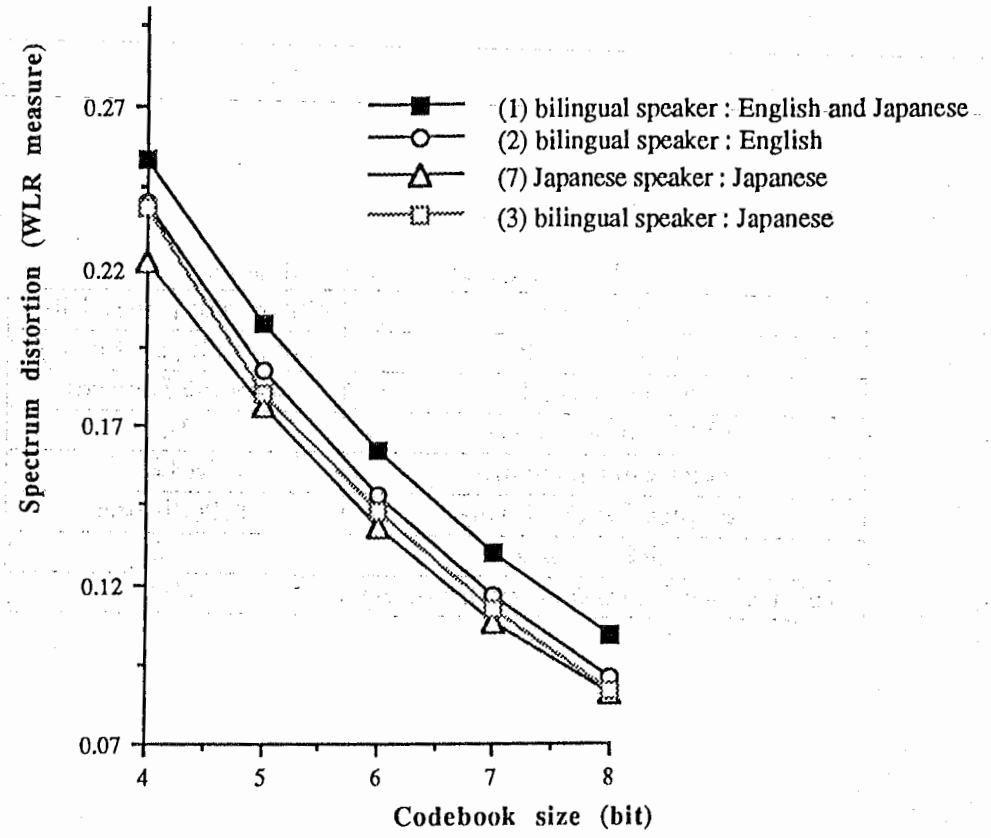


Fig.3 Spectrum distortion for various codebooks

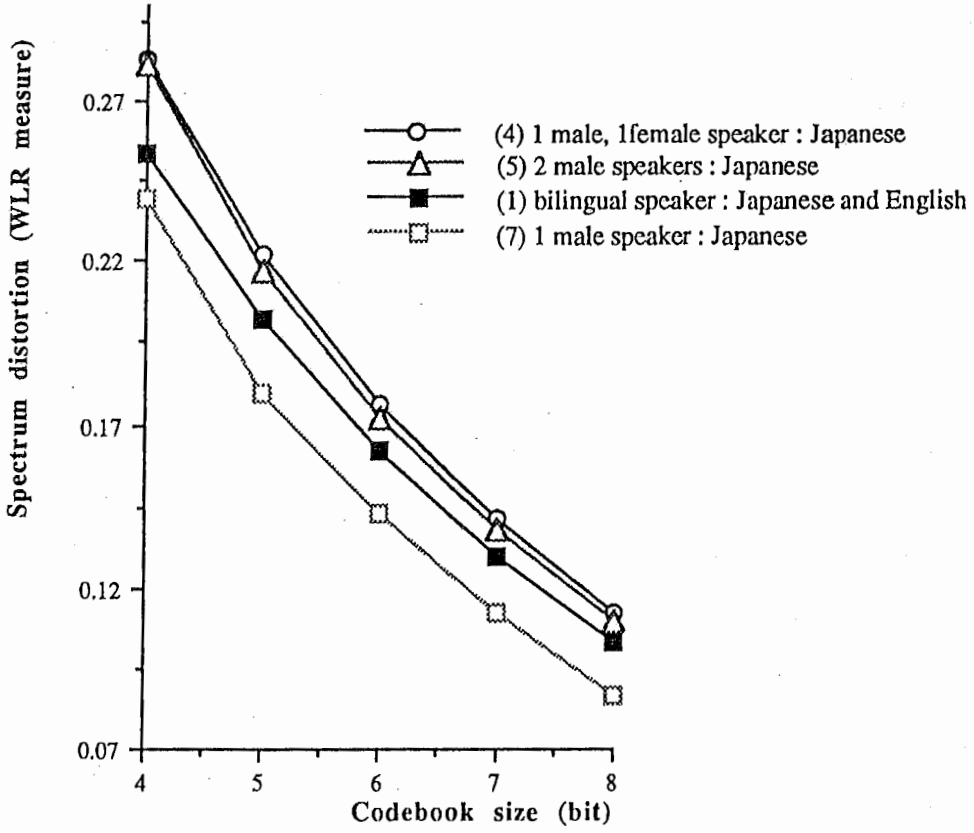


Fig.4 Spectrum distortion for various codebooks

Table 2 Kullback's divergence

Speaker pair	Kullback's divergence
English vs. Japanese	1.21
male speaker vs. female speaker	8.59
male speaker1 vs. male speaker2	4.80
word set1 vs. word set2	0.21

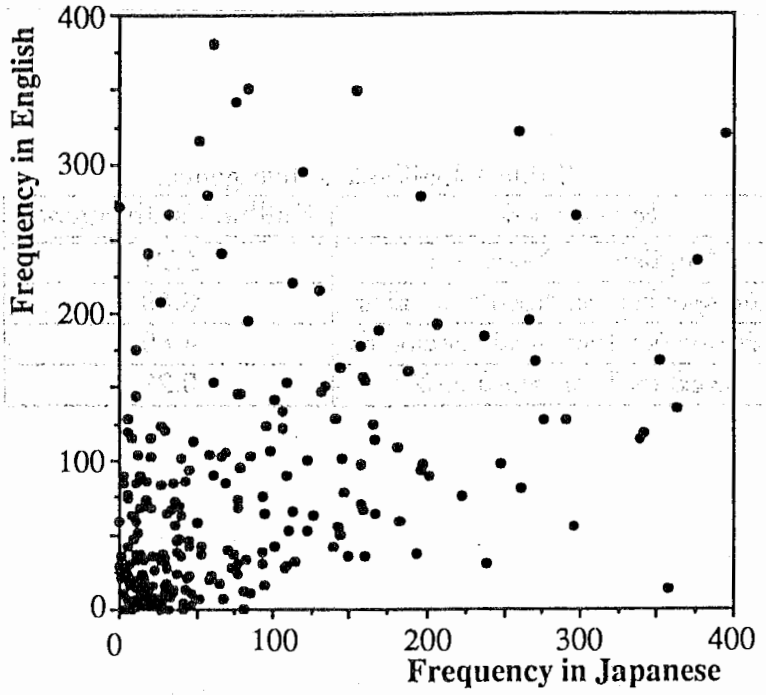


Fig.5 Frequency of codewords in Japanese and English

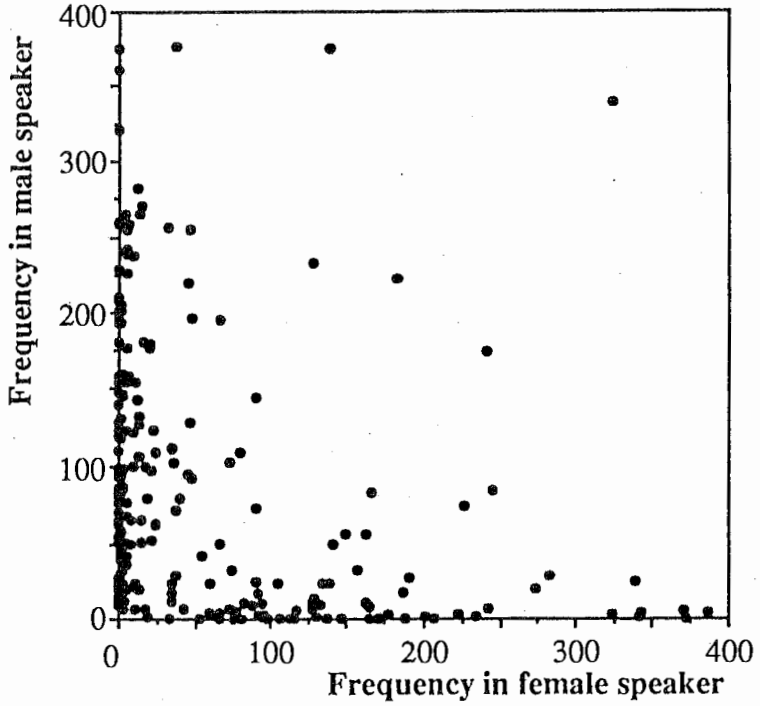


Fig.6 Frequency of codewords in male and female speakers

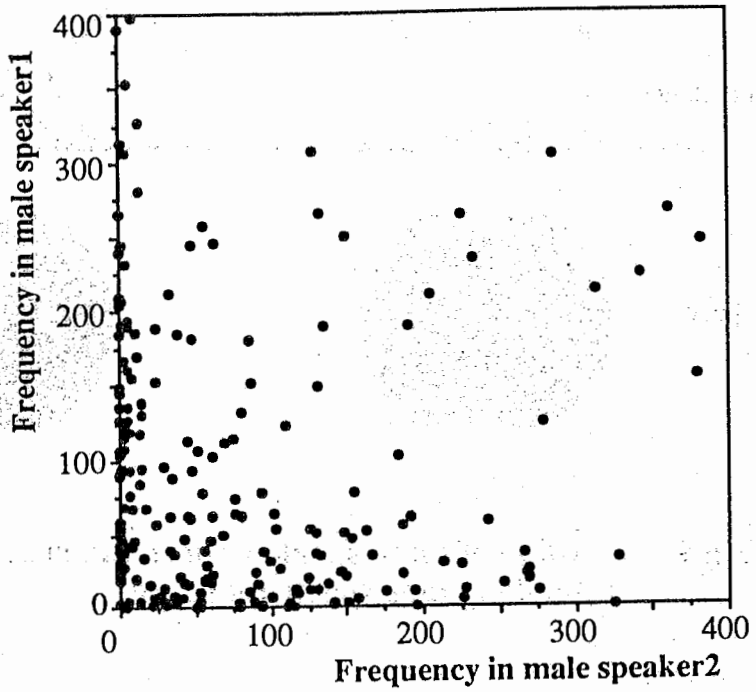


Fig.7 Frequency of codewords in different male speakers

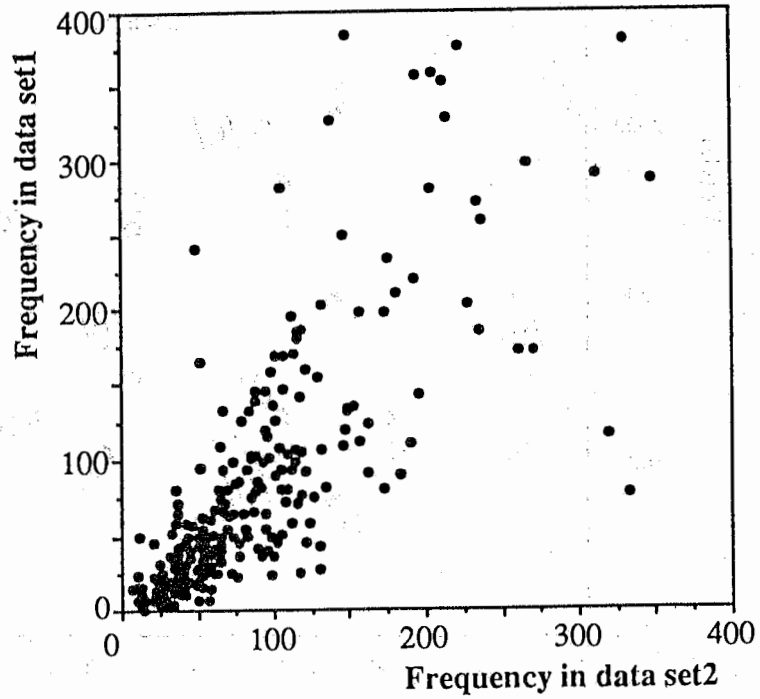


Fig.8 Frequency of codewords in the same male speaker

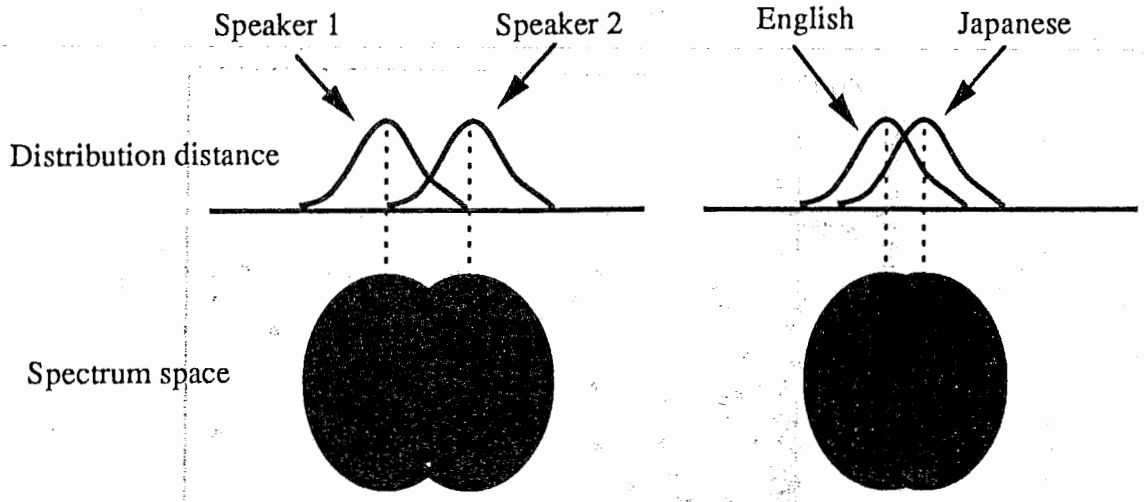


Fig.9 An outline of inter-speaker spectrum space vs. inter-language spectrum space

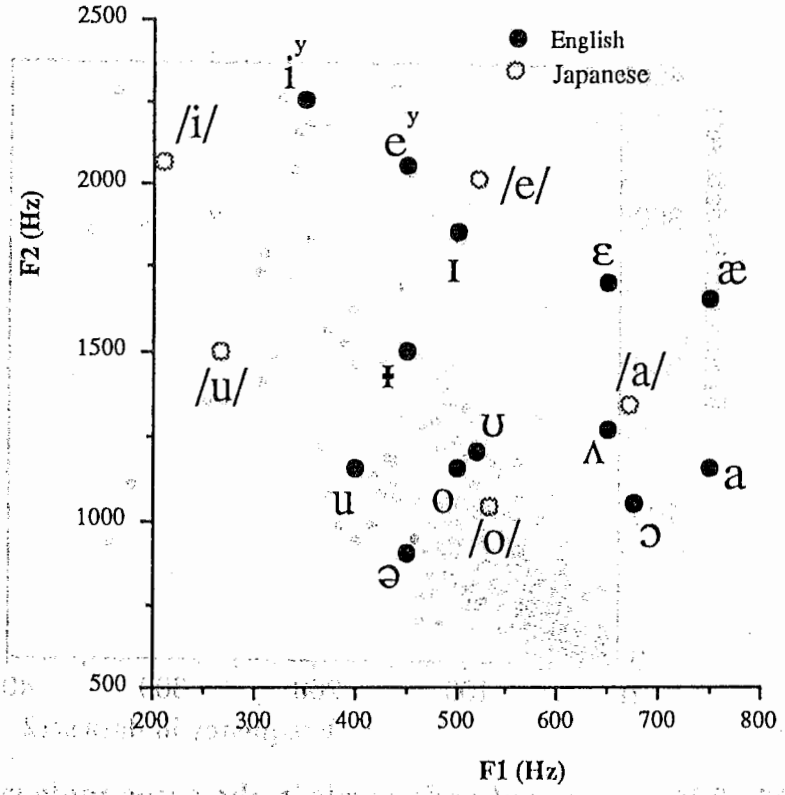
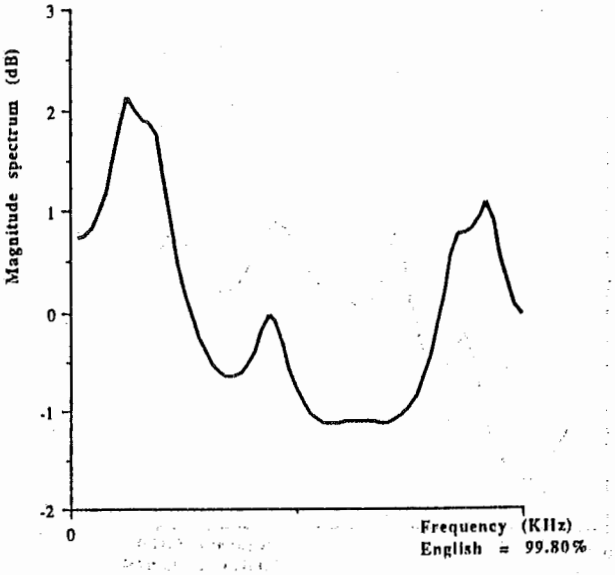


Fig.10 F1 and F2 frequency in English and Japanese



LPC spctrum envelope

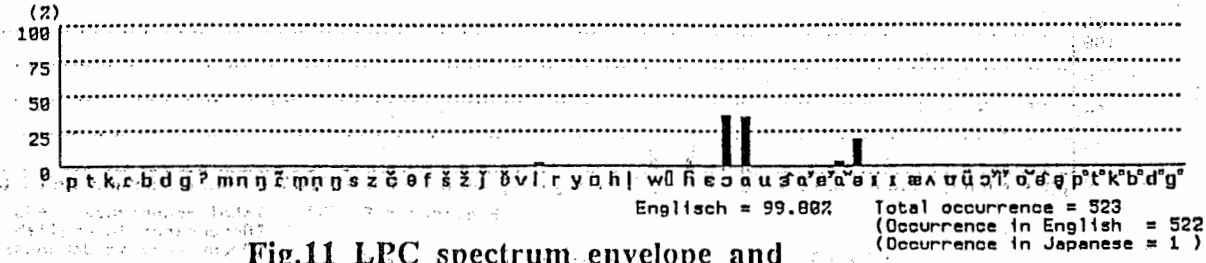
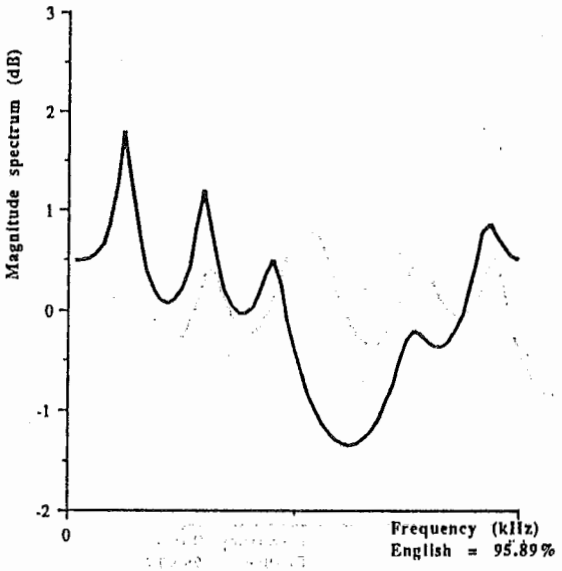


Fig.11 LPC spectrum envelope and phoneme histogram of a codeword



LPC spectrum envelope

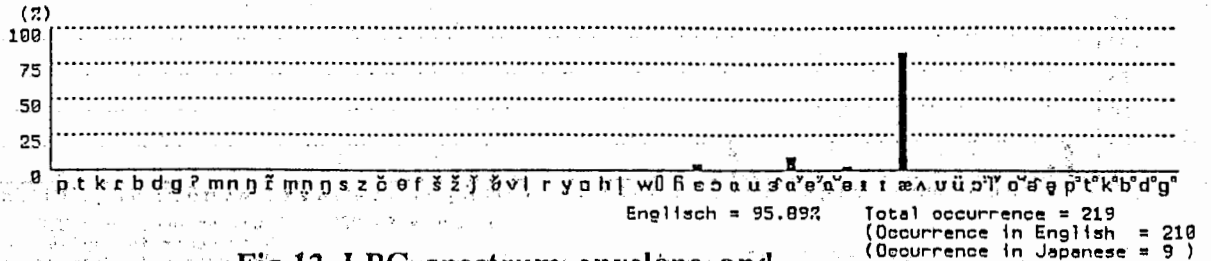
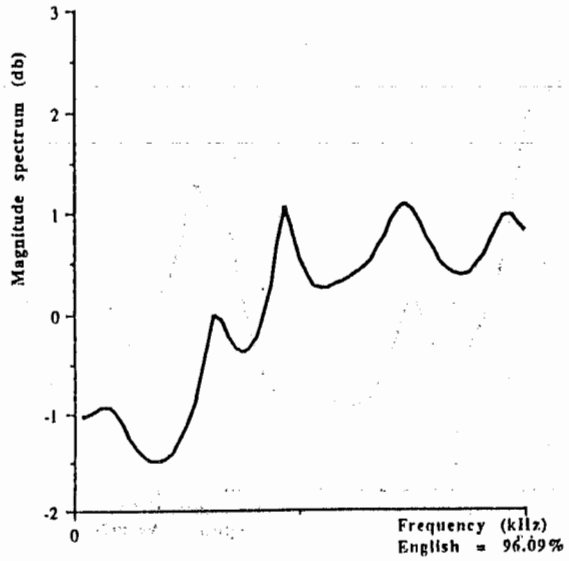


Fig.12 LPC spectrum envelope and phoneme histogram of a codeword



LPC spectrum envelope

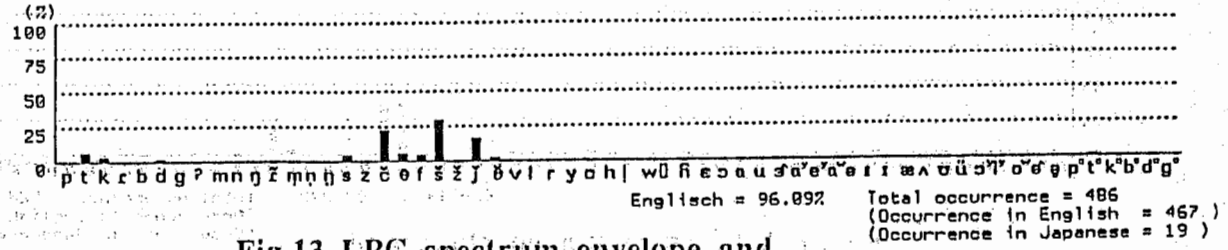
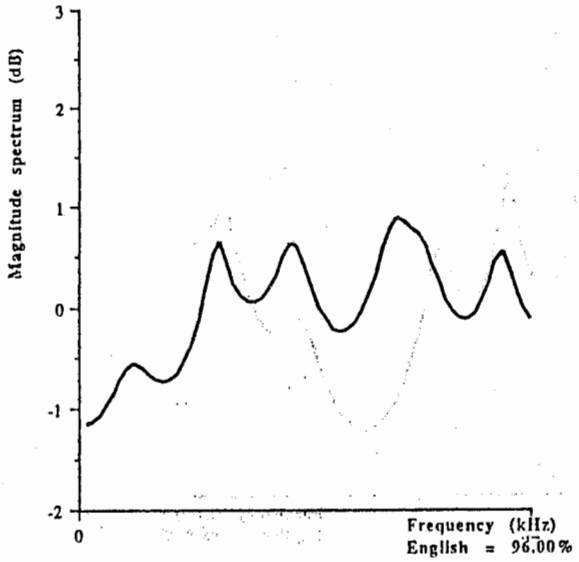


Fig.13 LPC spectrum envelope and phoneme histogram of a codeword



LPC spectrum envelope

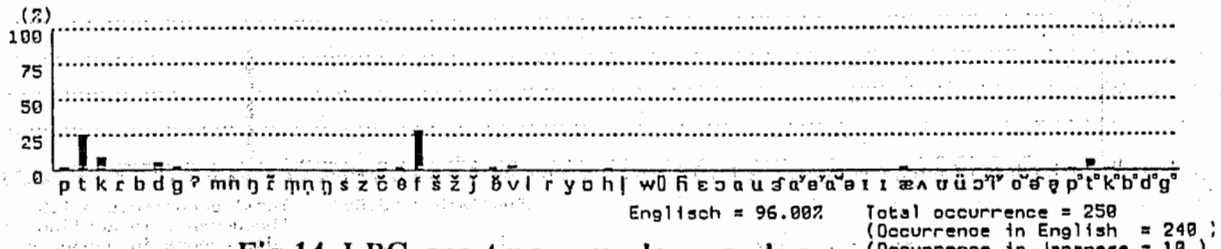


Fig.14 LPC spectrum envelope and phoneme histogram of a codeword

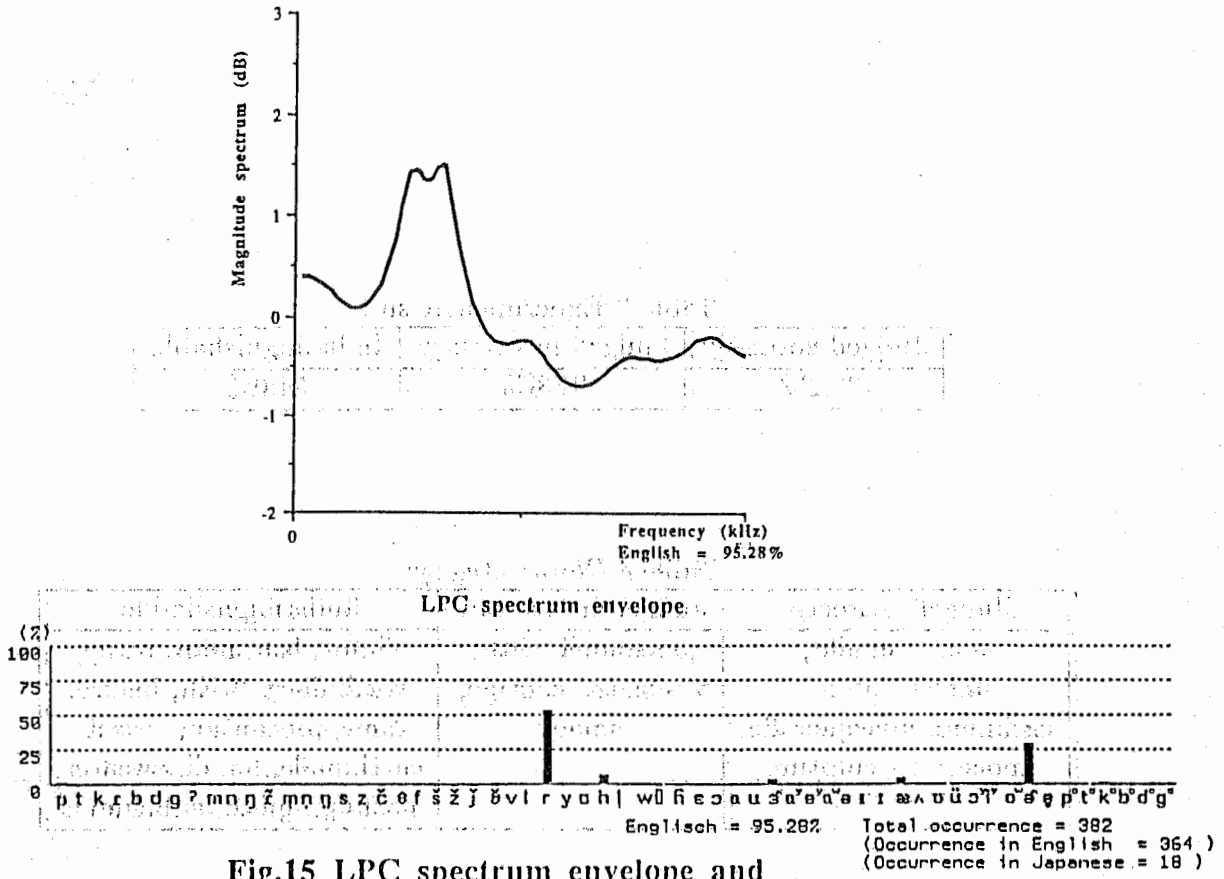


Fig.15 LPC spectrum envelope and phoneme histogram of a codeword

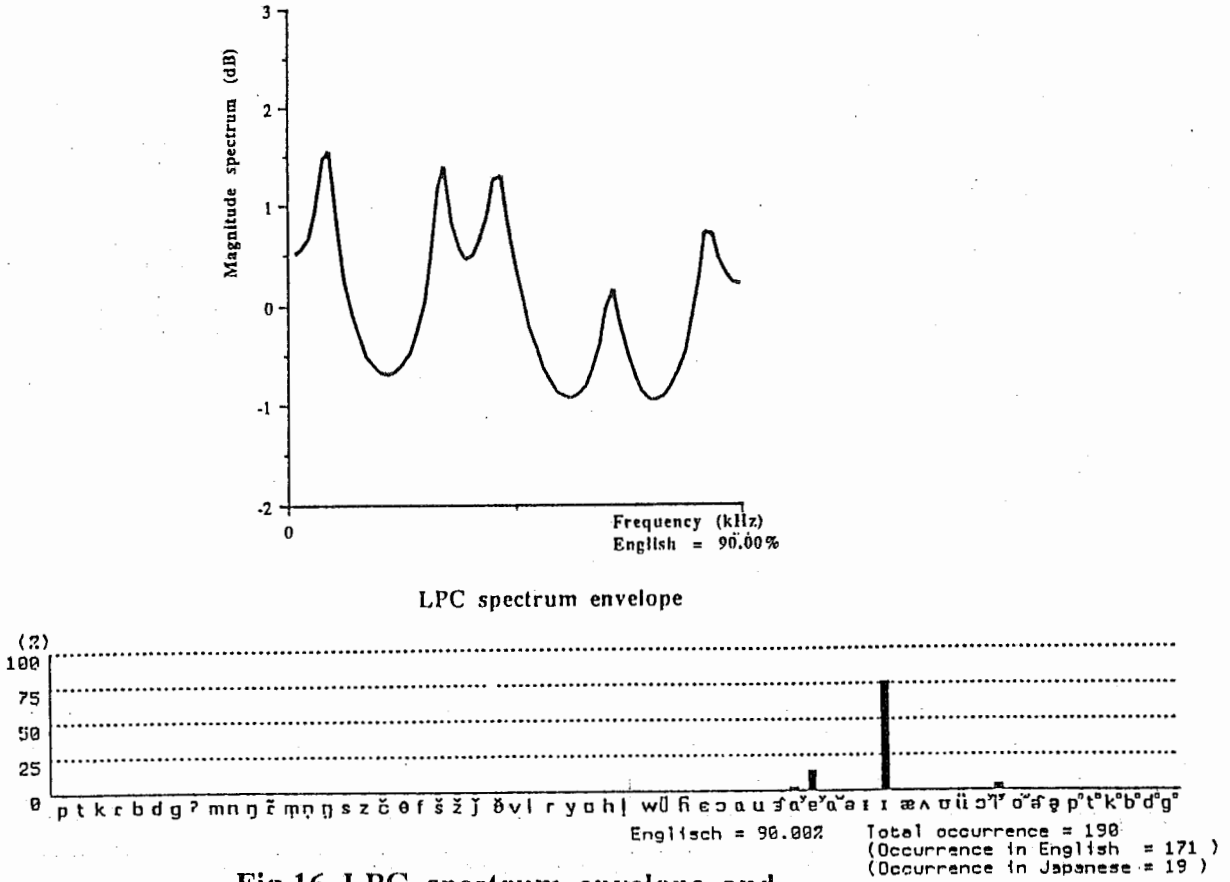


Fig.16 LPC spectrum envelope and phoneme histogram of a codeword

Table 3 Experiment result

Judged correctly	Judged incorrectly	Indistinguishable
27.2%	18.8%	54.0%

Table 4 Word category

Judged correctly	Judged incorrectly	Indistinguishable
noise, should, finger, outer, cashmere, masquerade, moisture, sculpture	personnel, with, zoologist, corsage, money	victor, fish, noteworthy, vocabulary, Irish, before, they, precaution, sweet, earthquake, hand, sweater, nothing, quite, ambiguous

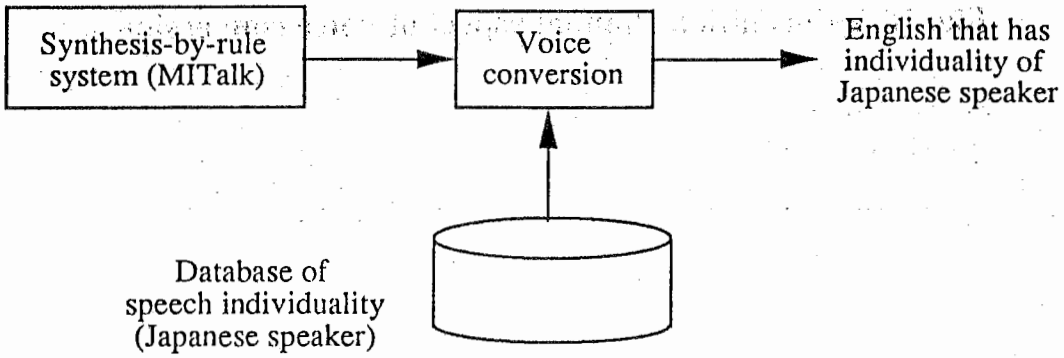


Fig.17 Cross-language voice conversion model

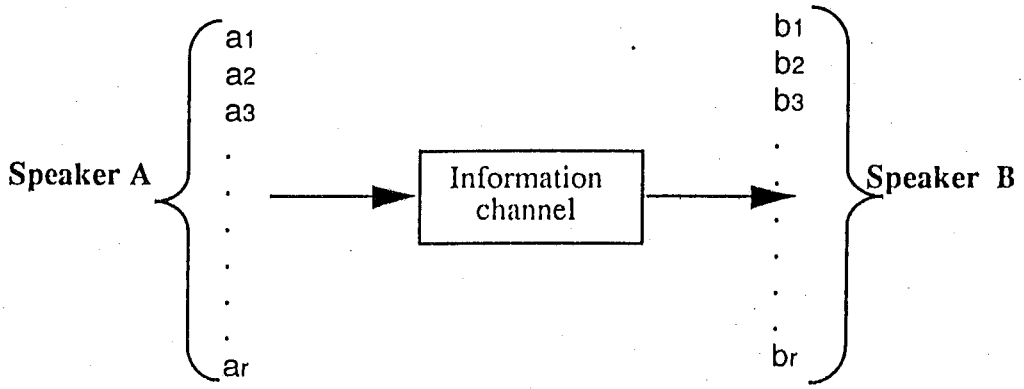


Fig. 18 Information channel aspect of voice conversion

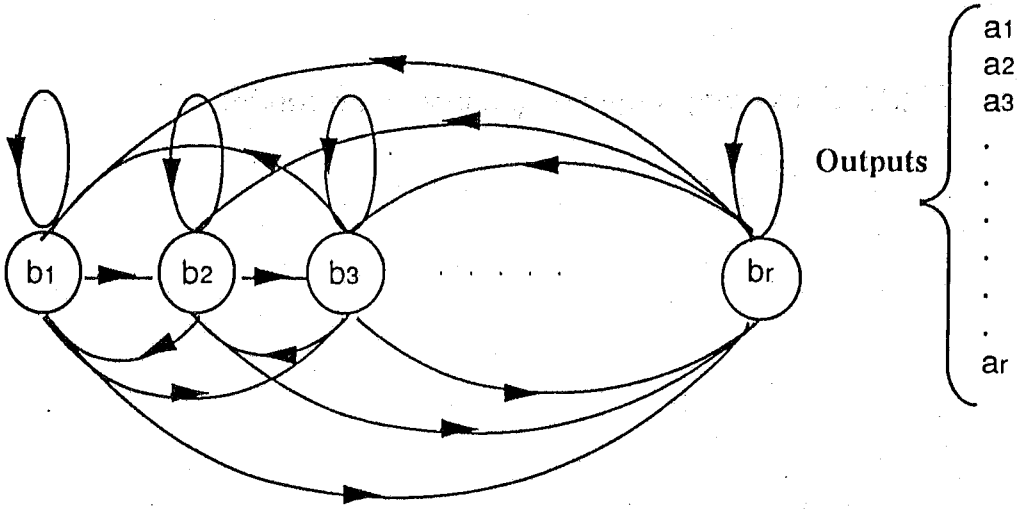


Fig.19 Speaker Markov model

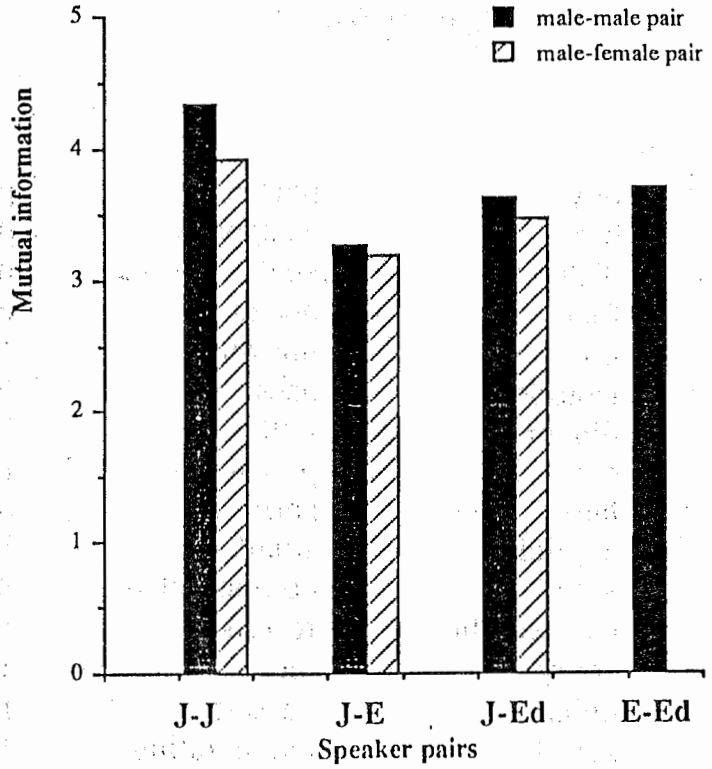


Fig.20 Mutual information for speaker pairs

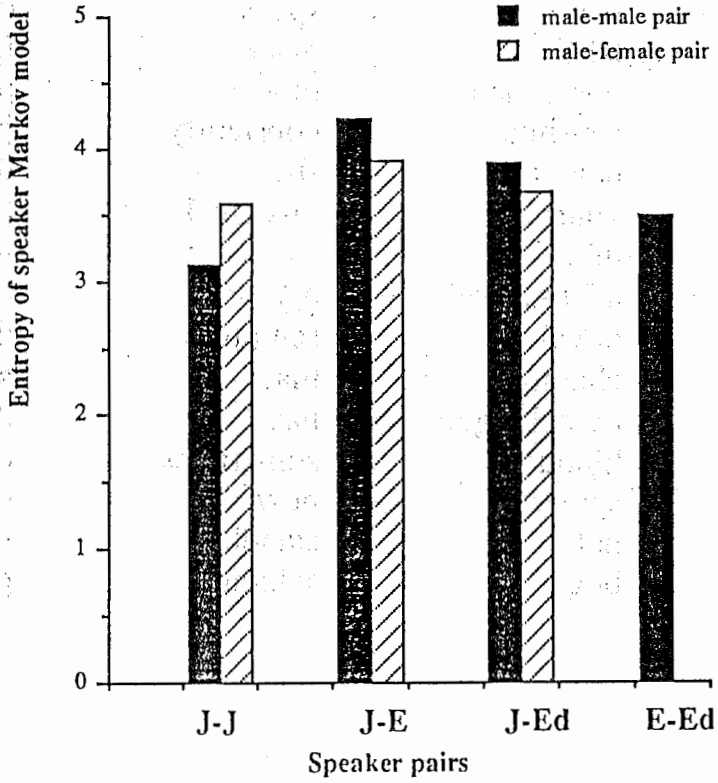


Fig.21 Entropy of speaker Markov model for speaker pairs

Appendix A

hat	buy	before	goose
clams	hear	teaspoons	michael
real	dish	heat	masquerade
welcome	fish	nectar	endurance
but	no	unevenly	sleeping
encouraged	isotopes	uses	teach
plow	dinner	will	buying
depicts	boy	precaution	price
smiles	tag	earn	save
need	explicitly	small	hand
farmland	victor	compounded	petticoats
reflects	overweight	fortune	difficult
plant	spray	pie	live
treats	bells	dessert	mediocrity
immediate	sweet	recuperating	lightbulbs
exam	documents	zoologist	woolen
catastrophic	annoying	cory	walking
entertaining	raisins	carefully	stayed
balls	street	steep	wealth
moon	two	speaker	controlled
anecdotal	lake	black	from
working	postponed	now	popular
sought	sunshine	constantly	judge
ahead	degrees	she	drunkard
screen	your	glistened	trouble
ambiguous	tube	as	remember
classrooms	affirmative	my	enough
motorists	down	sculpture	determination
items	chamber	that	moisture
every	overcharged	has	divorced
met	hindu	sometimes	cashmere
crooked	spider	how	corsage
muscular	not	standby	right
the	bog	seldom	greg

Appendix A (cont.)

farmyard	splurged	they	rob
alien	be	wasp	events
occasionally	several	new	rationalize
outer	purists	examples	occurs
tribes	previous	Bob	paper
do	priorities	sold	their
forest	may	path	have
breakfast	relaxed	Irish	nothing
radioactive	board	his	education
me	sweater	Gus	cameo
therapy	worship	boring	valley
spring	iguanas	in	I
traffic	zircons	noise	employment
finger	skirts	through	personnel
society	dry	barbed	child
aquatic	strong	made	elm
flower	played	offensive	at
potatoes	accusations	choosing	wash
Monday	please	if	sun
by	on	oasis	chipper
cash	should	techniques	shoulder
audition	house	an	along
big	system	allow	provoked
into	never	you	people
schooner	extra	standardized	each
too	prevented	cream	chives
dirty	straight	greatly	all
broke	hit	competition	voyage
about	afternoon	diane	pair
goes	many	greasing	up
objects	gently	wild	nightly
out	gives	overalls	agency
with	flew	refurbishing	seeking
near	cheese	saw	often

Appendix A (cont.)

simple
does
youngsters
juice
gas
noteworthy
capable
vocabulary
jaguars
one
of
living
blues
ambled
get
thinker
emphasized
high
must
variety
go
fawn
think
could
were
are
those
attitude
gallon
quite
screw
sound
three
is
woman
to
money
icicles
only
drift
her
alligators
way
brother
geological
was
for
number
when
had
apology
earthquake
large
antagonistic
a