Japanese Speaker-Independent Homonyms Speech Recognition

Jin'ichi Murakami Haseo Hotta
Dept. of Information and Knowledge Eng., Tottori University,
4-101 Koyama-Minami, Tottori 680-8550, Japan

Abstract

Japanese has homonyms such as "$\overline{ha}\underline{shi}$ "(箸 (Chopsticks)) and " $\underline{ha}\overline{shi}$ "(橋 (Bridge)). Word speech recognition has been studied for a long time, but homonym speech recognition in Japanese has not been studied. In this paper, we studied speaker-independent homonym speech recognition. For homonym speech recognition, pitch extraction has been normally used to estimate a pitch frequency. . However, we did not use pitch extraction in our study. Instead, we used an accent model that was a phoneme label with more length, mora position, accent type and accent high or low. It means that we used the effect of pitch on formant. The result of the experiments were that 89% accuracy was obtained by using MFCC, full covariance HMM, and the accent model.

1. Introduction

Japanese has homonyms like "箸" ($\overline{ha}\underline{shi}$, [chopsticks]) and "橋" ($\underline{ha}\overline{shi}$ [bridge]) [1]. These words have the same syllables but different accents. However, normal speech recognition uses formants and not prosodies [7]. So, homonym speech recognition in Japanese has not been studied[6][8].

In Chinese, the difference in the accent (tone) creates different meaning of a word. It is called "four-tone" or "tone sandhi". Thus, many prosody studies on speech recognition Chinese[2],[3] have been conducted. These research used both MFCC[1] and pitch frequency. MFCC indicates a formant structure, and pitch frequency indicates a part of prosody. However, reliably estimating pitch frequency has been very difficult. Double pitch and half pitch often estimated. Also vowels have pitch, but voiceless consonants do not.

In this study, we used the effect of pitch on formants, and did not directly use the extraction of pitch frequency. More specifically, we used an accent model based on the phoneme with word mora length, word mora position, the type of accent, and accent high or low.

Using this model, we studied speaker-independent homonym speech recognition. We also used a pair set of homonyms for an evaluation. In accent model, the number of syllables in HMM is too much. So we used semi-continuous HMM[5] in this study. We also used MFCC and FBANK[1] for the acoustic parameters.

[1] " " is Japanese kanji kana expression. []is English meaning

2. Accent Model and Accent Triphone Model

In this section, we describe the accent model. The model indicates the phoneme label with the word mora length and word mora position and the type of accent and accent high or low added. This accent model has vowels and nasal and double consonants. And the normal consonant does not have these labels.

More specifically, we labeled labeled vowels as well as nasal and double consonants with seven digit numbers. The first pair of numbers indicates the mora length for a word. The second pair indicates the word's mora position. The third pair indicates the word's accent type. The final number indicates the accent at the mora position. It is expressed using 0 for low and 1 for high. Fig. 1 shows an example of the models.

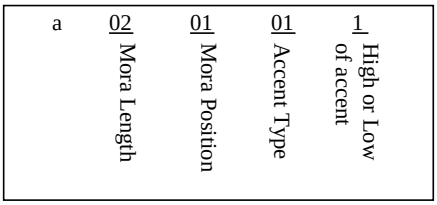

Figure 1. Label for Accent Model

This accent model is a context-independent accent model. In this paper, we also use an accent triphone model that is a context-dependent accent model. Table1 shows an example of the accent, accent triphone, and triphone models. Example word is "$\overline{a}\underline{ki}$". This word of Japanese kanji expression is "秋", and English expression is "Autumn". $\overline{a}$ indicates that the accent of "a" is high. $\underline{ki}$ indicates that the accent of "ki" is low. In this table, $+$ shows the after context dependent phoneme, and $-$ shows the before context dependent phoneme.

Table 1. Example of Labels

Word: a k i { “秋” [Autumn] }

phoneme model	a	k	i
accent model	a0201011	k	i0202010
triphone model	a+k	a−k+i	k−i
accent triphone model	a0201011 +k	a0201011− k +i0202010	k− i0202010

3. Homonym Recognition Experiments

3.1. Training Data and Test Data

We used an ATR A-set database. This database has 5240 words spoken by each of ten male and ten female speakers. Speakers were professional and voiced very clearly. For the training data, we used nine speakers, and odd number words. That is, we used 2620 x 9 words for training, and other one speaker was kept for testing.

For the test data, we used homonym data for a speech database. To survey the word accent, we used the “NHK Japanese Accent Dictionary”[4]. The ATR A-set database had 31 pairs totaling 62 words. However, the speech data had different accents. Thus, we used correctly accented words in this database. As a result, we used 11 pairs of homonyms (i.e., 22 words). Table 2 shows the test homonyms data. In this table, $\overline{syllable}$ indicates that the accent of the “ syllable ” is high and $\underline{syllable}$ indicates that the accent of the “ syllable ” is low. “ ” is Japanese kanji expression and [] is English meaning.

Table 2. Evaluation Data (Pairs of Homonyms)

1.	$\underline{i\overline{ru}}$ “居る” [stay]	$\overline{\underline{i}ru}$ “射る” [shoot]
2.	$\underline{ka\overline{eru}}$ “代える” [change]	$\overline{\underline{ka}eru}$ “返る” [reverse]
3.	$\underline{ka\overline{keru}}$ “欠ける” [missing]	$\overline{\underline{ka}keru}$ “駆ける” [run]
4.	$\underline{ki\overline{geN}}$ “機嫌” [mood]	$\overline{\underline{ki}geN}$ “起源” [origin]
5.	$\underline{ko\overline{ukai}}$ “公開” [public]	$\overline{\underline{ko}ukai}$ “航海” [voyage]
6.	$\underline{o\overline{ku}}$ “置く”[carry]	$\overline{\underline{o}ku}$ “億” [A hundred millions]
7.	$\underline{shi\overline{mei}}$ “指名” [nominate]	$\overline{\underline{shi}mei}$ “氏名” [full name]
8.	$\underline{ta\overline{bi}}$ “度” [at a time]	$\overline{\underline{ta}bi}$ “足袋” [Japanese socks]
9.	$\underline{to\overline{ku}}$ “徳” [virtue]	$\overline{\underline{to}ku}$ “解く” [solve]
10.	$\underline{tu\overline{keru}}$ “付ける” [attach]	$\overline{\underline{tu}keru}$ “漬ける” [steep]
11.	$\underline{yo\overline{ru}}$ “因る” [cause]	$\overline{\underline{yo}ru}$ “夜” [night]

3.2. Experimental Conditions

We conducted an experiment with three male speakers and three female speakers. We used the HTK tool kit [1] and FBANK and MFCC in these experiments. Also, we used full covariance HMM and diagonal covariance HMM. MFCC and FBANK have the same number of Gaussian densities.

Table 3 shows acoustic analysis parameters and the parameters of HMM. The experimental conditions are also shown in table4.

Table 3. Acoustic Parameters

record frequency	16 kHz
window length	25 ms
frame period	10 ms
Number of analyses (MFCC)	12 order MFCC + Δ 12 order MFCC + log power + Δlog power
Number of analyses (FBANK)	24 order FBANK + Δ24 order FBANK + log power + Δ log power

Table 4. HMM Parameters

HMM model	3 loop 4 state semi continuous densities
# stream	3
# Gaussian densities of state (Diagonal)	MFCC 1024 + ΔMFCC 1024 + log power 64 + Δ log power 64
# Gaussian densities of state (Full)	MFCC 128 + ΔMFCC 128 + log power 16 + Δ log power 16

3.3. Flowchart of Making Accent Model and Accent Triphone Model

The initial HMM is very important to training. And data spareness for accent model and accent triphone model is a serious problem. Thus, we made the initial accent model HMM from a phoneme model HMM, and the initial triphone model HMM was made from the phoneme model HMM . Also, the initial accent triphone model HMM was made from triphone models HMM. Also, to avoid the problem of data spareness for accent model and accent triphone model, we used semi continuous HMM[1].

Figure2 shows the flowchart for the accent model HMM and accent triphone model HMM.

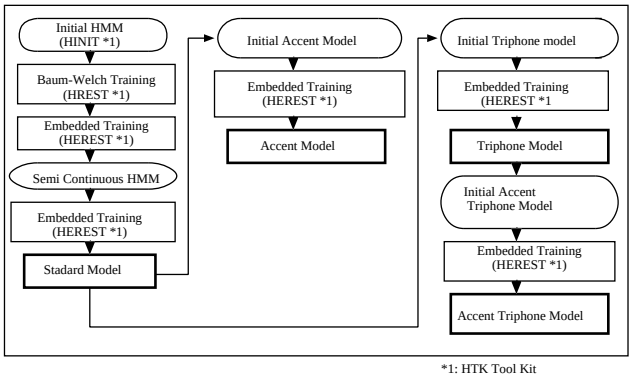

Figure 2. Flowchart of Making Accent Models HMM and Accent Triphone Models HMM

4. Results of Homonym Speech Recognition

Tables 5 show the results of speaker independent homonym speech recognition. In this table, "MAU", "MMY", and "MNM" indicate a male speaker, and "FAF", "FMS", and "FTK" indicate a female speaker. "Ave.(Male)" indicates the average of male speakers (MAU MMY MNM). "Ave.(Female)" indicates the average of female speakers (FAF FMS FTK). "Ave.(Total)" indicates the average of all speakers. Table5 shows the results of the error rate using "MFCC and Diagonal Covariance HMM", "MFCC and Full covariance HMM", "FBANK and Diagonal Covariance HMM", and "FBANK and Full covariance HMM".

The following results were obtained in these experiments.

1. Best Model

   The maximum average homonym recognition rate (89%) was obtained for the accent triphone model and MFCC and full covariance HMM (table5). However, the results differed between male speakers and female speakers.

2. Males vs. Females

   Female speakers had a higher recognition rate than male speakers. Male speakers had a higher MFCC recognition rate than FBANK. Female speakers had the opposite trend. The maximum recognition rate of male speakers was 92% with the accent triphone model and MFCC and full covariance HMM (table 5). The maximum recognition rate of female speakers was 94% with the accent triphone model and FBANK and full covariance HMM (table 5) .

3. MFCC vs. FBANK

   The average MFCC recognition rate was slightly higher than the average FBANK recognition rate. MFCC was effective with male speakers, and FBANK was effective with female speakers.

Table 5. Results (Error Rate)

Speaker	Accent model	Accent triphone model
MFCC, Diagonal		
MAU	27%(6/22)	18%(4/22)
MMY	18%(4/22)	27%(6/22)
MNM	36%(8/22)	27%(6/22)
FAF	23%(5/22)	18%(4/22)
FMS	9%(2/22)	0%(0/22)
FTK	6%(6/22)	27%(6/22)
Ave. (Male)	27%(18/66)	24%(16/66)
Ave. (Female)	20%(13/66)	15%(10/66)
Ave. (Total)	23%(31/132)	20%(26/132)
FBANK, Diagonal		
MAU	23%(5/22)	27%(6/22)
MMY	23%(5/22)	27%(6/22)
MNM	41%(9/22)	32%(7/22)
FAF	23%(5/22)	23%(5/22)
FMS	5%(1/22)	0%(0/22)
FTK	32%(7/22)	18%(4/22)
Ave. (Male)	29%(19/66)	29%(19/66)
Ave. (Female)	20%(13/66)	14%(9/66)
Ave. (Total)	24%(32/132)	21%(28/132)
MFCC, Full		
MAU	14%(3/22)	5%(1/22)
MMY	23%(5/22)	5%(1/22)
MNM	32%(7/22)	14%(3/22)
FAF	5%(1/22)	5%(1/22)
FMS	9%(2/22)	9%(2/22)
FTK	27%(6/22)	27%(6/22)
Ave. (Male)	23%(15/66)	8%(5/66)
Ave. (Female)	14%(9/66)	14%(9/66)
Ave. (Total)	18%(24/132)	11%(14/132)
MFANK, Full		
MAU	18%(4/22)	14%(3/22)
MMY	27%(6/22)	32%(7/22)
MNM	45%(10/22)	32%(7/22)
FAF	0%(0/22)	9%(2/22)
FMS	5%(1/22)	0%(0/22)
FTK	14%(3/22)	9%(2/22)
Ave. (Male)	30%(20/66)	26%(17/66)
Ave. (Female)	6%(4/66)	6%(4/66)
Ave. (Total)	18%(24/132)	16%(21/132)

4. Accent Model vs. Accent Triphone Model

   In most cases, the accent triphone model was better than the accent model. However, the difference was small between the two. It was large only with MFCC and full covariance HMM. The error rate improved 23% to 8%, whereas the improvement was not very large in other experiments.

5. Discussion

5.1. Analysis of homonym recognition error

Across experiments, errors for homonym recognition were 2 mora high low and 3 mora low high high words. Table 6 shows an example of the errors for 2 mora homonyms. As shown in this table, these homonyms are easy errors for people.

Table 6. Example Errors for Homonym Recognition (2 mora)

Output	Correct
$\underline{o}\overline{ku}$ “置く” [carry]	$\overline{o}\underline{ku}$ “億” [A hundred millions]
$\underline{yo}\overline{ru}$ “因る” [cause]	$\overline{yo}\underline{ru}$ “夜” [night]
$\underline{i}\overline{ru}$ “居る” [stay]	$\overline{i}\underline{ru}$ “射る” [shoot]

5.2. Comparison of FBANK and MFCC

MFCC was more effective than FBANK for speaker-independent homonym recognition in many experiments. However, among female speakers, FBANK was more effective than MFCC in many experiments. FBANK has prosodies and formants, information on both prosodies and formants, while MFCC has only formant information. However, the prosodies affect the formants. Thus, homonym speech recognition is possible even with MFCC. However, FBANK seems better overall than MFCC for homonym speech recognition.

This hypothesis for speaker independent speech recognition holds true on female speech but incorrect on male speech.

5.3. Comparison of Males and Females

There are no differences in accent components of relative f0 and intensity between the male and female groups. Normally, female speakers generally have higher pitch frequency. It makes difficult to separate formant and pitch. Thus, female speakers are worse than male speakers at normal speech recognition.

However the opposite results were obtained with homonyms speech recognition. The error rate of homonym speech recognition is lower for female speakers than male speakers. We think that the change in female speakers' pitch frequency is larger than the change in male speakers' pitch frequency, thereby providing support for this conclusion.

5.4. Comparison of proposed method and other models

We must compare of proposed method and other models. As pitch extraction is the most important point in the paper. So we will have a data by the proposed systems with separate pitch extractor.

6. Conclusions

In this study, we surveyed the recognition rates of Japanese speaker-independent homonym speech. To recognize the homonyms, we created an accent model and an accent triphone model. An accent model had a phoneme label with word mora length and word mora position and the type of accents and accent high or low. An accent triphone model had a triphone label with word mora length and word mora position and the type of accents and accent high or low. Also, we did not use pitch extraction. For acoustic parameters, we used MFCC and FBANK.

Using these models and parameters, we studied the homonym speech recognition rates. And we obtained the following results.

1. Using accent triphone models, MFCC, and full covariance HMM, we obtained 89% homonym word accuracy.
2. The MFCC produced higher average recognition rates than FBANK, meaning that it was generally more effective. However, MFCC was better than FBANK for male speakers, and FBANK was better than MFCC for female speakers.
3. Much difference was evident in the recognition rates of the speakers.

In the future, we will use FBANK because this parameter is effective for speaker dependent recognition and for female speakers. Or we will use other parameters like LDC. And we will use discriminative training for HMM.

References

[1] Steve Young, etc, “HTK Version 3.2 reference manual”, Cambridge University, 2002.

[2] Yi-hao, K., Lin-shan, L., “Feature Analysis for Emotion Recognition from Mandarin Speech Considering the Special Characteristics of Chinese Language”, InterSpeech 2006: 1814-1817, 2006.

[3] Dau-Cheng Lyu, Min-Siong Liang, Yuang-Chin Chiang, Chun-Nan Hsu, Ren-Yuan Lyu, “Large Vocabulary Taiwanese (Min-nan) Speech Recognition Using Tone Features and Statistical Pronunciation Modeling”, Eurospeech 2003, 1861-1864, 2003.

[4] “NHK 日本語発話アクセント辞典新版 (Japanese Accent Dictionary)”, NHK 出版, 1998. ISBN4-14-011112-7. (In Japanese)

[5] Huang, X. D., Ariki, Y., and Jack, M. A., “ Hidden Markov models for speech recognition”, EDinburg University Express, ISDN 0-7486-0162-7 , 1990.

[6] Jinichi Murakami, Terou Araki and Satoru Ikehara, “ Information of Pose and Accent in Japanese Speech ”, 1988 Autumn Meeting Acoustical Society of Japan, No. 3-3-11, pp. 89–90, 1988.

[7] Lee, K.-F., Automatic Speech Recognition: The Development of the SPHINX SYSTEM, Kluwer Academic Publishers, Boston, 1989.

[8] Keikichi Hirose, Frederic Gendrin and Nobuaki Minematsu, “ A Pronunciation Training System for Japanese Lexical Accents with Corrective Feedback in Learner's Voice”, Eurospeech 2003, 3149-3152, 2003.