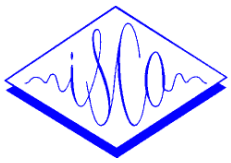

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

ISOLATED WORD RECOGNITION USING PITCH PATTERN INFORMATION

Satoshi Takahashi, Sho-ichi Matsunaga, and Shigeki Sagayama

NTT Human Interface Laboratories
3-9-11, Midori-cho, Musashino-shi, Tokyo, 180 JAPAN

ABSTRACT

This paper describes a new technique for isolated word recognition that uses both pitch information and spectral information. Words with similar phonetic features tend to be misrecognized in conventional methods which use only spectral information, even if their phonemes are accented differently. Many phonetically-similar Japanese words are classified by pitch patterns. This paper introduces a measure of the pitch pattern distance. A pitch pattern template is produced by averaging pitch patterns obtained from a set of words which have the same accent pattern. A measure for word recognition is proposed, based on a combination of the pitch pattern distance and the phonetic likelihood. Speaker-dependent word recognition experiments were carried out using 216 Japanese words uttered by five male and five female speakers. The proposed measure reduces the recognition error rate by 40% compared with the conventional phonetic likelihood.

1. INTRODUCTION

Speech information contains both spectral information and prosodic information. Prosodic information includes pitch, duration, and power information. In multiple vector quantization [1] or power-spectrum vector quantization [2], the power pattern is effectively used to improve recognition accuracy. However, a recognition method using the global pitch patterns of Japanese words has never been reported.

When humans listen to speech, prosodic information plays a very important role as well as spectral information. It is very difficult for humans to recognize the word without the accent. Some words have quite different meanings depending on the accent, which is the only clue for distinguishing the meaning.

Conventional automatic word recognition systems only use spectral information. Therefore, they tend to confuse phonetically similar words such as "kyūryō" and "kyūgyō". The word "kyūryō" is always uttered with an accent at the beginning of the word. On the other hand, "kyūgyō" has no accent. Pitch pattern information is very useful for distinguishing these two words. A word recognition system with a large vocabulary will have many such word-pairs. The combination of spectral information and pitch information is more important in a large-vocabulary system.

This paper describes a new technique of word recognition that uses both pitch pattern information and spectral information. Word utterance recognition is carried out based on both the acoustic likelihood and the accentual likelihood.

2. SYSTEM OVERVIEW

A block diagram of the system is shown in Figure 1. The system has two major process flows. The left flow is a phoneme recognition part based on a Hidden Markov Model (HMM). The other side is for pitch pattern recognition. Pitch patterns are classified based on template matching using Dynamic Time Warping (DTW). The system includes phonetic HMMs, pitch pattern templates, and a word dictionary. The word dictionary includes a phoneme sequence and an accent type for each word.

The input speech is analyzed by LPC analysis to obtain LPC parameters. The LPC parameters are vector-quantized

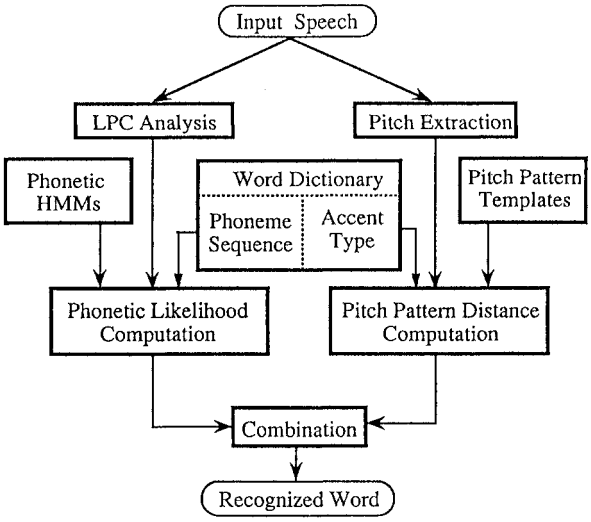

Fig. 1 Word Recognition System Using Pitch Pattern Information

and represented by a code number. The system refers to a phoneme sequence of each word and calculates the phonetic likelihood of the input speech by concatenating phonetic HMMs as a word template.

The pitch pattern recognition is executed in parallel with the phoneme recognition. After pitch extraction, the pitch is smoothed and normalized. Since Japanese words are classified by accented syllable position, several universal pitch pattern templates are made by averaging word pitch patterns pronounced with a typical accent. The pitch pattern distance between the input pitch pattern and the pitch pattern template is calculated using DTW. The system combines the pitch pattern distance and the HMM phonetic likelihood to recognize a word utterance.

3. PITCH PATTERN TEMPLATES

3.1 Accent types of Japanese words

Two pitch levels, high and low, are enough to classify the pitch patterns of Japanese words. A Japanese word of N syllables can be uttered in N different accent patterns. Figure 2 shows an example for a 4-syllable word, where a circle represents a syllable and the arrow indicates the descent position of the pitch. The type of pitch pattern, called the accent type, is defined by the number of syllables from the beginning of the word to the accented syllable. Generally, Japanese words start at a low pitch and go up to a higher pitch for the second syllable. This high pitch continues until the accented syllable, after which the pitch descends to a lower pitch toward the next syllable. The descent of the pitch value plays a very important role in the perception of the accent. There are two exceptions: type 0 and 1. In type 0, the pitch never descends because there are no accented syllable. In type 1, the pitch starts high and descends just after the first syllable.

3.2 Pitch pattern templates

The pitch pattern templates for each accent type are generated in the following way. Pitch extraction uses the lag window method, where the original spectrum is divided by the smoothed spectrum obtained through a lag window in the auto-correlation domain [3]. The maximum value of the inverse FFT of the divided spectrum is denoted ρ max. The pitch value of the input data is calculated from the position of ρ max. In addition, the value of ρ max is useful in judging whether the obtained pitch value is reliable or not. In general, pitch values with a small ρ max, usual in the unvoiced part, are considered unreliable. An unreliable pitch value is eliminated beforehand in order to make the pitch pattern contour clear. The threshold for acceptance of ρ max, ρτ, is determined from the distributions of ρ max in the unvoiced sound and in the voiced sound. The pitch values are smoothed using a five-point median smoother and converted to logarithms. All patterns are linearly warped to adjust their length to the average duration (800 ms) of the training words. The pitch values in each word are normalized by the average pitch value. Pitch patterns which have the same accent type are averaged to make a template. In type 1, the pitch pattern of a word whose first phoneme is a voiced consonant is slightly different from the pitch pattern of other words [4]. Therefore, type 1 has two pitch pattern templates. Figure 3 shows examples of the pitch pattern templates.

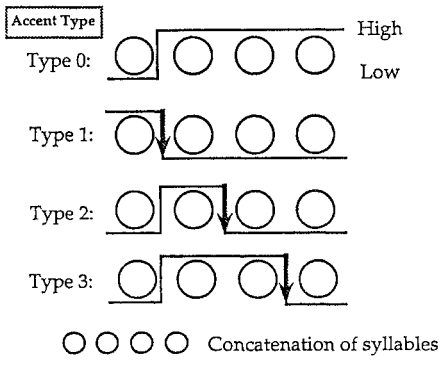

Fig. 2 Pitch patterns of 4-syllable Japanese words

4. PITCH PATTERN MATCHING

Pitch patterns are distinguished by the position of the descent from high-pitched syllable to low-pitched syllable. Since the whole pitch pattern of a word shifts back and forth along the time axis according to the duration of the first syllable, we applied endpoint-free DTW to calculate the distance between the input speech and the template. The distance between pI, the log pitch value of the i-th frame in the input speech, and tJ, that of the j-th frame in the template, is calculated as follows.

$$\text{if} \quad \rho\,\max(i) \geq \rho\tau \qquad d(i,j) = |\,pi - tj\,|^2 \qquad (1)$$

$$\text{if} \quad \rho\,\max(i) < \rho\tau \qquad d(i,j) = (\text{not calculated})$$

where $\rho\,\max(i)$: ρ max of the i-th frame in the input speech

$\rho\tau$: threshold of ρ max for deciding whether pitch value is reliable or not.

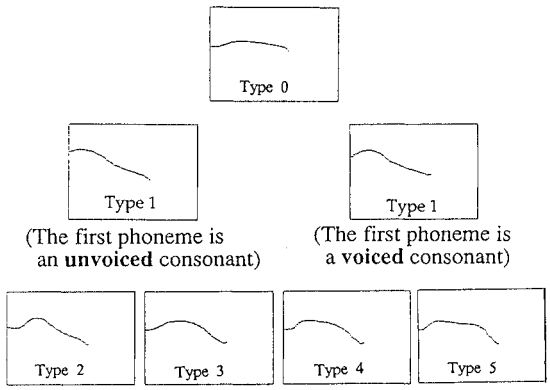

Fig. 3 Pitch Pattern Templates

The produced pitch pattern has a gap at an unvoiced part. It is possible to interpolate the pitch value during the unvoiced part from the pitch of neighboring voiced parts. However, at the transition between the voiced part and the unvoiced part, pitch extraction errors often occur and thus pitch values are unstable. So, the cumulative distance is calculated using only the voiced frame. Pitch pattern distance is obtained by the cumulative distance normalized by the number of values considered in the distance calculation.

5. ACCENT TYPE DISCRIMINATION EXPERIMENT

As a preliminary experiment, an accent type discrimination experiment was carried out to check the system's ability to distinguish pitch patterns using two sets of 216 phonetically-balanced words uttered by five male and five female speakers. One set was used for making pitch pattern templates and the other was used for testing. A speech researcher listened carefully to all recorded words to define the accent types for the word dictionary.

5.1 Experiment using male speaker data

Six pitch pattern templates, types 0 to 5, were made from one data set of each speaker. Table 1 shows the confusion matrix for the male speaker, showing that it is difficult to distinguish types 2 to 5. According to the definition of the accent type, the difference between types 2 to 5 is the number of syllables from the first syllable to the accented syllable. However, the length of the high-pitched syllables varies depending on the phonemes composing the word, and it makes difficult to discriminate between those patterns. Type 0 also tends to be misrecognized as type 5. The difference between these two patterns is whether the pitch descends at the end of the word or not. In long words of type 0, the pitch pattern tends to show a downtrend because of the long utterance, and the final frequency reaches that of type 5. In this case, the perceived pitch level is still high but the physical pitch level is actually low at the end of the word. Thus, the physical pattern and the perceived pattern are not always the same.

We regarded the four pitch patterns, types 2, 3, 4, and 5, as one category. This gave a total of three categories; type 0, type 1, and type N (types 2 to 5). The discrimination rates for three categories are summarized in Table 2. The shaded boxes indicate the discrimination rate, where the average discrimination rate is 90.5%.

5.2 Experiment using female speaker data

Table 3 shows the results for the female speaker. Although pitch extraction for a female voice is more difficult than for a male voice, the discrimination ability is similar in our experiment. Table 3 shows that there is more confusion in distinguishing type 0 and type N than the case of male speaker. The difference between types 0 and N is heavily dependent on the end of the pitch contour, where the pitch value cannot be extracted accurately because of the weak utterance.

Table 1 Confusion matrix of the accent type discrimination experiment (male speakers)

Test Sample \ Judgement	0	1	2	3	4	5
0	506	1	0	1	5	45
1	5	143	6	0	5	2
2	0	9	56	12	4	2
3	8	2	35	77	29	14
4	2	0	9	16	25	17
5	4	0	1	3	5	31

Table 2 Confusion matrix after categorizing types 2 to 5 as type N (speaker-**dependent** condition, **male** speakers)

Test Sample \ Judgement	0	1	N
0	90.7	0.2	9.1
1	3.1	88.8	8.1
N	3.9	3.0	93.1

N : type 2 - type 5

Table 3 Confusion matrix after categorizing types 2 to 5 as type N (speaker-**dependent** condition, **female** speakers)

Test Sample \ Judgement	0	1	N
0	85.0	0.2	14.8
1	1.8	91.8	6.4
N	10.5	2.6	86.9

N : type 2 - type 5

6. WORD RECOGNITION EXPERIMENT USING PITCH PATTERN

This new technique is evaluated by speaker-dependent word recognition experiments using the system shown in Figure 1. The speech data is the same as that used in the accent type recognition experiment. One set of words is used for making pitch pattern templates and phonetic HMMs. The other set is used for testing.

6.1 Phoneme recognition

The feature vector has 34 dimensions: 16 cepstrum coefficients, 16 delta cepstrum coefficients, power, and delta power. A discrete HMM with a single code book is used. All learning data is phonetically labeled by hand. A total of 25 phonetic HMMs are generated. The word likelihood is calculated using the Viterbi algorithm, by referring to the phoneme sequences in the dictionary.

6.2 Combination

Parameter α is introduced to combine the pitch pattern distance and the phonetic log likelihood, to obtain the total score (Sc), as shown in Equation (2). The pitch pattern distance is subtracted so that a larger total score is better.

$$Sc = (1.0 - \alpha) * DPL - \alpha * DAC , \quad (2)$$

where DPL: phonetic log likelihood
DAC: pitch pattern distance.

6.3 Results

Figure 4 shows the change in the number of misrecognized words as a function of α. This result includes both learning and testing data. When $\alpha = 0$, i.e. using only phonetic likelihood, 26 words were misrecognized. Of these misrecognized words, 13 words had a different accent type from the correct word's accent type. It is possible to recover these misrecognition by introducing the pitch pattern distance. When α was set to 0.1, 11 misrecognized words were recovered. No side effects were observed.

Table 4 shows the final word recognition results of the testing data uttered by five male and five female speakers. The number of misrecognized words was reduced from 24 to 14 for the male speakers, and from 11 to 5 for the female speakers. The proposed measure incorporating the phonetic likelihood and the pitch pattern distance reduced the error rate by 40% compared with the case using only spectral information.

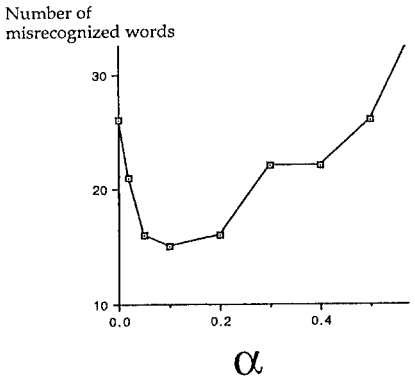

Fig. 4 Number of misrecognized words as a function of parameter α

Table 4 Word recognition results for testing data

	5 male speakers [1080 words]	5 female speakers [1080 words]
HMM phoneme likelihood	24 (2.22%)	11 (1.02%)
HMM phoneme likelihood + Pitch pattern distance	14 (1.30%)	5 (0.46%)

Number of errors (error rate)

7. CONCLUSION

A new isolated word recognition technique using both pitch patterns and spectrum patterns is proposed. This technique can distinguish phonetically-similar words, if their phonemes are accented differently.

A method to produce the pitch pattern templates and the pitch pattern distance to measure the difference of two pitch patterns were introduced. This paper also proposed a measure combining the phonetic likelihood and the pitch pattern distance. Its effectiveness was proven through speaker-dependent word recognition experiments.

ACKNOWLEDGEMENT

The authors wish to thank Dr. Sadaoki Furui, Dr. Kiyohiro Shikano, and Dr. Hirokazu Satoh, for their continuous support of this work. We also wish to thank all the members of the Speech and Acoustic Laboratory for their helpful discussion and constant encouragement.

REFERENCES

[1] K-F. Lee, H. W. Hon, "Large-Vocabulary Speaker-Independent Continuous Speech Recognition Using HMM", Proc. ICASSP88, S3.7, 1988.

[2] K. Aikawa, K. Shikano, "Spoken Word Recognition Using Vector Quantization in Power-Spectrum Vector Space", The Trans. of the Institute of Electronics, Information, and Communication Engineers of Japan, Vol. J68-D, No.3, pp1-7, 1985, in Japanese.

[3] S. Sagayama, S. Furui, "Pitch Extraction Method Using Lag Window", The Institute of Electronics, Information, and Communication Engineers of Japan, Autumn National Convention Record, 1235, 1978, in Japanese.

[4] H. Satoh, "Analysis of Fundamental Frequency Characteristics Related to Phonemes", The Acoustic Society of Japan, Fall Meeting Proc.,2-3-18, pp259-260, 1989, in Japanese.

[5] S. Takahashi, S. Matsunaga, S. Sagayama, "Isolated Word Recognition Using Pattern Information", Institute of Electronics, Information, and Communication Engineers of Japan Technical Report, S90-17, pp65-72, 1990, in Japanese.