A New Word Spotting Framework Using Hough Transform of Distance Matrix Images

Hiroyuki Nishi, Yoshimasa Kimura, Ren Iguchi, *

Abstract—**In this report, a new word spotting framework using Hough transform of distance matrix images is proposed. The brightness of an intersectional pixel in the distance matrix image expresses the distance value between an input and a standard pattern. If a standard pattern word is included in the input utterance, the straight line area is observed in the distance matrix image. If the voting number for the corresponding Hough transform parameters is more than the threshold value, it is decided that the standard pattern word exists in the input utterance. This new framework was evaluated by the recognition accuracy, the correct/incorrect words rejection rate and the false alarm.**

Keywords: Speech recognition, Word spotting, Distance matrix image, Hough transform

1 Introduction

Word spotting is a useful method for practical spoken dialog systems, because the system can recognize a user's ideas, even when unnecessary words are uttered before or after the keywords. However, both the start and end points of the word need to be given in conventional word spotting methods[1][2][3][4][5], for example continuous Dynamic Programming or the Hidden Markov Model. Therefore, it is necessary that a flexible and effective language model is provided and the score of the whole utterance can be calculated.

This report proposes a method of discovering a key word among the entire utterance by expressing input speech as image information and observing a line in the image using the Hough Transform[6][7].

2 Distance Matrix Images

Figure 1 shows how to obtain the distance matrix image. The vertical axis is the frame number of the standard pattern speech and the horizontal axis is that of the input speech. The intersection data is calculated as the cepstrum distance between the corresponding standard pattern frame and the input speech frame. The distance data is converted into the pixel brightness. If the distance is 0 (namely, where the input data frame is the same as the standard pattern frame), the brightness scale is 255 (a white pixel), whereas if the distance is infinite, the brightness scale is 0 (a black pixel).

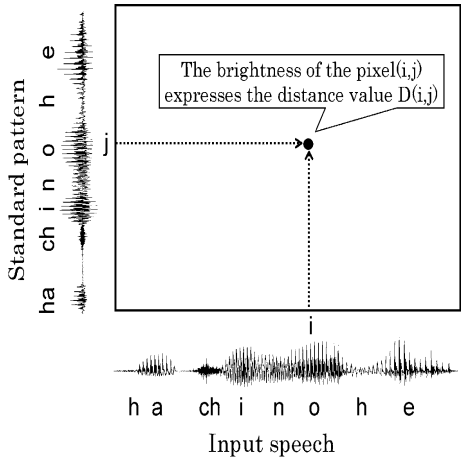

Figure 1: Explanation of a distance matrix image

The typical distance matrix image where the standard pattern is included in the input speech is shown in Fig. 2. In this case, a white line area indicated by the dotted circle can be observed because similar characteristic parameters as the standard pattern word are input.

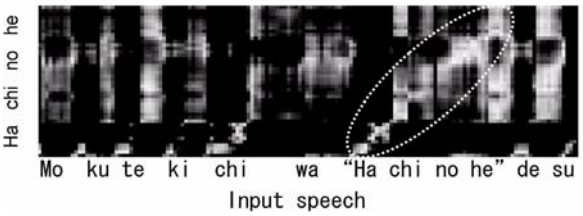

Figure 2: Sample of a distance matrix image (when standard pattern is included)

Conversely, in a distance matrix image (shown by Fig. 3) where the standard pattern is excluded, no white line area is observed.

Therefore, if such a white line area is detected using image

*SOJO university, Faculty of Computer and Information Sciences, Kumamoto Japan 860-0082 Tel/Fax: +81-96-326-3659/3000
Email: nishi@cis.sojo-u.ac.jp

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

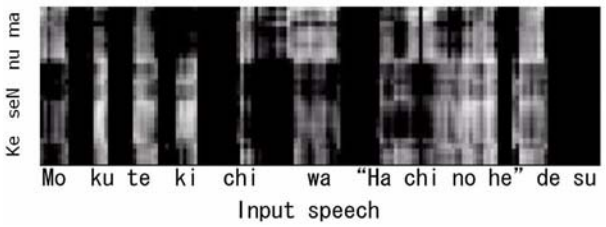

Figure 3: Sample of a distance matrix image (when standard pattern is excluded)

processing technology, word spotting becomes possible without determining the start and end points of the word.

In order to detect a line in a graphical image, the minimum mean square method and Hough transform are well known as strong and easy methods. Each has its own feature.

Minimum mean square method Because one straight line is estimated from the entire image data, data not related to the straight line also influence straight line parameters.

Hough transform As only data related to the estimated straight line are used, there is no influence from data of other areas.

With the above conditions in mind, Hough transform is clearly advantageous.

In the next section, the principle of Hough transform that detects a straight line from the graphical image is shown.

3 Hough transform

The principle of Hough transform is shown in Fig. 4.

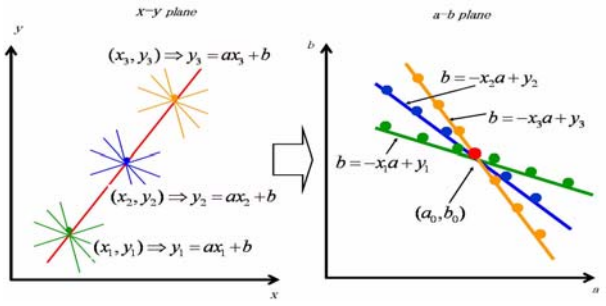

Figure 4: Principle of Hough Transform

To simplify the explanation, we assume there are three points in the x-y plane. Initially, the data (x_1, y_1) is examined.

Many straight lines can be drawn in the surroundings of point (x_1, y_1). These lines are expressed as $y_1 = ax_1 + b$.

The difference of the radial lines is indicated by the difference of the parameter set (a, b).

Each radial line has its own line parameter. Therefore, the corresponding data set, namely, the value of (a, b) can be mapped to the a-b plane. This operation is called a voting. As a result of one voting, one dot is described in the a-b plane.

Voting by the radial lines around (x_1, y_1), dots in the a-b plane line up on a straight line, and the relationship between a and b is given as $b = -x_1a + y_1$.

Using the same method, voting can be conducted related to data (x_2, y_2) and (x_3, y_3).

Consequently, the coordinate values (a_0, b_0) having acquired the most votes become the parameters for the straight line (gradient and intercept).

4 Binarizing distance matrix images

Before the voting processes of Hough transform, the distance matrix images must be binarized.

Only the white dots are the objects of the voting process, which means that the input frames are close to the corresponding standard pattern frames.

Therefore, the calculation volume increases in tandem with the number of white dots. However, if the amount of white dots is insufficient, the clear straight line cannot be described.

Conversely, when the amount of white dots is greater than the appropriate number, no straight line area appears. The typical binarized distance matrix image where a standard pattern word is included in the input speech with the appropriate threshold is shown in Fig. 5.

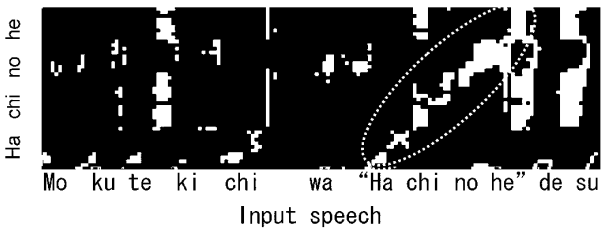

Figure 5: Sample of a binarized distance matrix image with the appropriate threshold

Figure 6 shows the binarized distance matrix image with the extremely small threshold.

In this case, the number of white dots is insufficient.

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Figure 6: Sample of a binarized distance matrix image with the small threshold

The third figure shows the binarized distance matrix image with the excessively large threshold.(Fig. 7.)

In this case, the number of white dots is overly excessive and no straight line area can be observed.

Figure 7: Sample of a binarized distance matrix image with the large threshold

5 Procedure of word spotting using the new method

5.1 Training procedure

The training procedure is the same as that of the conventional word speech recognition system. The power and cepstrum data of each word are analyzed and stored into the database.

5.2 Word spotting procedure

The word spotting process involves finding one of the standard patterns with the shortest distance to a part of the input speech.

To do so, the following three types of processing are necessary:

1. The system calculates the distance matrix image of one standard pattern and the input, and finds the straight line using Hough transform. The voting number of the obtained line is defined as the score of the standard pattern.

2. Among all the standard patterns, the one whose score is the highest is selected as the final candidate.

3. If the score of the final candidate exceeds the threshold, the system outputs the candidate as the word spotting result. If the score of the candidate is lower than the threshold, the system judges the input as including no standard patterns, i.e. out of vocabulary.

Word spotting procedure using the Hough transform is described in Fig. 8. "j" is the standard pattern number. The loop in Fig. 8 is for one standard pattern that contains the first item of the above three processes.

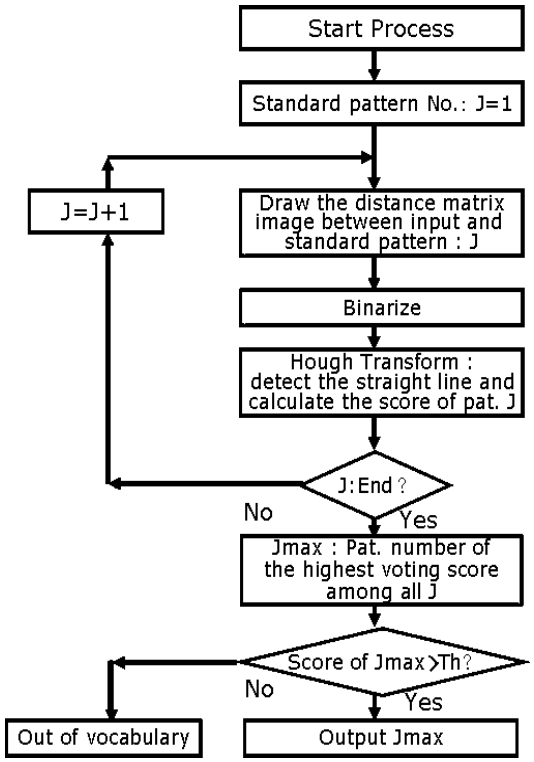

Figure 8: Procedure of word spotting using the Hough Transform

Selecting Jmax is for the second process of the three items.

The final judgment in Fig. 8 is the process that decides whether the first candidate is the correct word or "out of vocabulary".

6 Experiments

6.1 Experimental conditions

A new word spotting method was evaluated using the names of 100 Japanese cities standardized by JAIDA. The experimental conditions are shown in Table 1.

To set the same conditions, input sentences are constructed as "Mokutekichi ha", a keyword (city name)

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Table 1: Conditions of experiment

Items	Experimental Conditions
Analysis conditions	Sampling Freq.: 22.05kHz Bits of one sample: 16b Frame length: 24ms Frame shift: 12ms
Analysis method	Pow, LPC Cep, MFCC
Standard pattern words	JEIDA 100 cities
Input sentences	"Mokutekichi ha [city name] desu" (The destination is [city name].)
Number of speaker	1 male
Evaluation items	· Recognition accuracy · Error rate · Correct word rejection · Incorrect word rejection · False alarm
Parameters	· Threshold of distance value to binarize the : Th_d distance matrix images · Threshold of voting number to adopt the first candidate as the recognition result : Th_v

and "desu". "Mokutekichi ha" means "The destination". "desu" means "is". Therefore, the meaning of the sentence is "The destination is (a city name)".

Our new word spotting method was evaluated using the following 5 items.

Recognition accuracy N_{ca}/N_{kw} : The rate by which the first candidate is correct and the result is adopted when one of the key words is included in the utterance.

Error rate N_{ia}/N_{kw} : The rate by which the first candidate is incorrect and the result is adopted when one of the key words is included in the utterance.

Correct word rejection N_{cr}/N_{kw} : The rate by which the first candidate is correct and the result is rejected when one of the key words is included in the utterance.

Incorrect word rejection N_{ir}/N_{kw} : The rate by which the first candidate is incorrect and the result is rejected when one of the key words is included in the utterance.

False alarm N_{fa}/N_{oov} : The rate by which the first candidate is adopted when no key word is included in the utterance.

N_{kw}: The number of utterances which include key words.
N_{oov}: The number of utterances which include no key words.
N_{ca}: The number of utterances in which the first recognition candidate is the correct key word and the voting number is more than the voting threshold Th_v.
N_{cr}: The number of utterances in which the first recognition candidate is the correct key word and the voting number is less than the voting threshold Th_v.
N_{ia}: The number of utterances in which the first recognition candidate is the incorrect word and the voting number is more than the voting threshold Th_v.
N_{ir}: The number of utterances in which the first recognition candidate is the incorrect word and the voting number is less than the voting threshold Th_v.
N_{fa}: The number of utterances in which the voting number of the first recognition candidate is more than the voting threshold Th_v.

6.2 Experimental result

Figure 9 shows the recognition accuracy when one of the key words is included in the utterance.

MFCC is better than the LPC cepstrum as well as other conventional methods. However, the recognition accuracy depends on the threshold to binarize.

Particularly, threshold dependency in the case of MFCC is greater than in the case of LPC. Therefore, careful study for deciding the threshold is necessary.

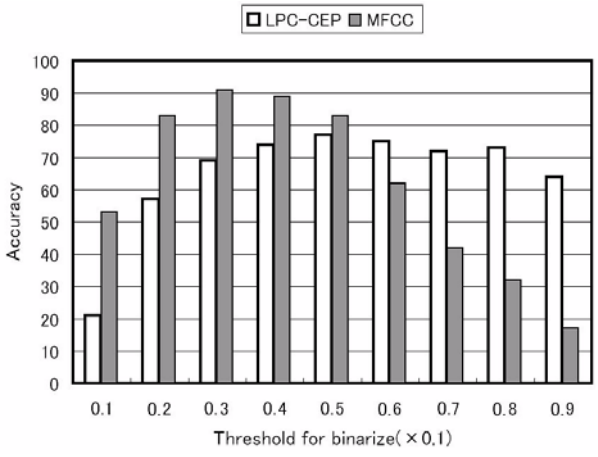

Figure 9: Recognition accuracy

The threshold to binarize influences the ratio of white/black pixels. Figure 10 shows the relationship between the threshold to binarize and the ratio of white pixels. Judging from Fig. 9, $(0.3-0.4)$ is appropriate for the situation of this experiment. On the other hand, looking at Fig. 10, the ratio of white pixels should be about 0.1 for the binarized threshold of $0.3-0.4$. Therefore,

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

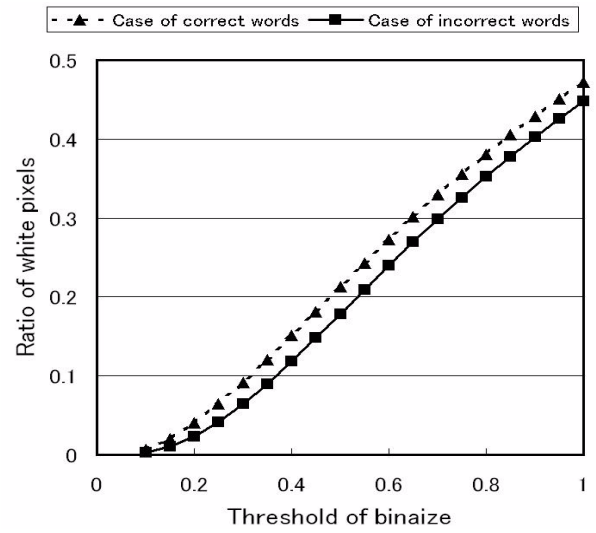

Figure 10: Relationship between the threshold and the ratio of white pixels

the threshold to binarize should be decided as the average white pixels ratio becomes about 10% in the distance matrix images for new appreciations or other situations.

Performance detail of the proposed method is shown in Fig. 11. This figure expresses multiple results.

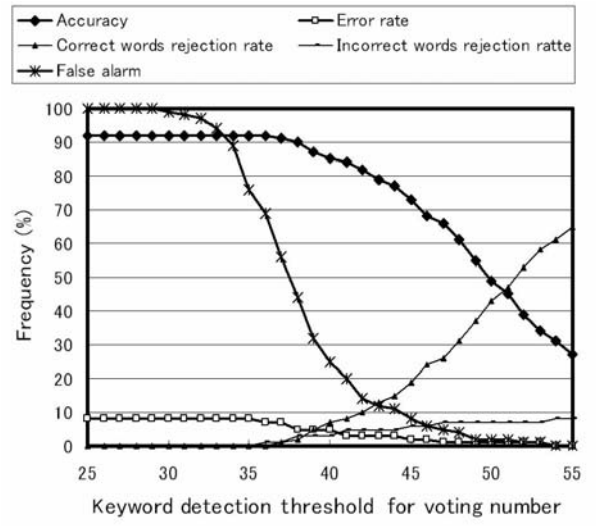

Figure 11: Performance detail of the proposed method

If no threshold of the voting number is given, the recognition accuracy is almost 90%. This accuracy is not as high as the word recognition system.

Figure 12 is a sample of an error.

The pronunciation of "bisai" in the input speech is similar to that of "hisai" in the standard pattern.

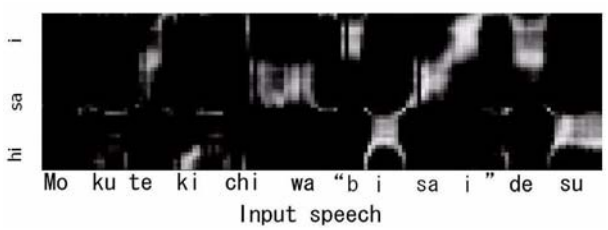

Figure 12: Example of recognition error(1) – Similar Pronunciation

Both Bisai and Hisai are names of cities in Japan. Only the first phoneme "b" and "h" differs.

Even though the intersection part of "b" and "h" in Fig. 12 is black and indicates that the distance is long, a high voting number is indicated by the other white straight line area.

Conversely, in the case of word spotting, other difficult situations obstruct the performance (Fig. 13).

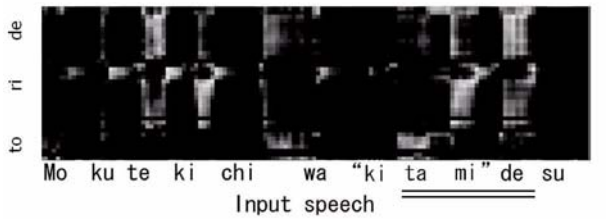

Figure 13: Example of recognition error(2) – False alarm

Figure 13 shows the error sample whose standard pattern is "toride" and the input speech is "Mokutekichi ha kitami desu".

In the figure, the straight line area covers "tami de" not "kitami". In other words, the combination of "tami", a part of the word "kitami", and "de" becomes "tamide", and mismatched ""toride".

In order to avoid this situation (False alarm), the threshold of the voting number to adopt the recognition result should be decided carefully.

Figure 11 shows that if the threshold of the voting number is selected as about 43, false alarm can be reduced to about 15%. At the same time, it is advantageous that the error rate decreases by 5%. However, in such condition we should note that the accuracy decreases to near 80%.

It is natural that the basic recognition accuracy should be further improved. However, it is more important that the appropriate threshold of the voting number must be selected according to the application's demand.

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

7 Conclusions

The new word spotting framework that uses the Hough transform of distance matrix images, where the brightness of the pixel in the intersection coordinates expresses the distance value between an input voice and a standard pattern, is proposed.

The availability of the new framework is confirmed by the word spotting experiment.

However, this study is just begun and a lot of themes have accumulated. We are planning a speaker independent and large vocabulary word spotting system using the Hough transform of distance matrix images.

8 Acknowledgement

We would like to thank Mr. Suehiro Nakamura, the vice president of Sojo University, for his helpful suggestions and discussions.

References

[1] NAKAGAWA S, HAUPTMANN A G, TOMITA M: "On quick word spotting techniques," Proc. ICASSP, No.Vol.3 Page.2311-2314 (1986)

[2] KAWABATA T, KOHDA M: "Word spotting method taking account of duration change characteristics for stable and transient parts of speech," Proc. ICASSP, Vol. 3 pp. 2307-2310 (1986)

[3] Kazuyo Tanaka, etc: "Constructing Spoken Document Processing Systems on a universal subphonetic segment domain," Proc of 9th Western Pacific Acoustics Conference(WESPAC IX 2006), Paper no.459(CD-ROM)

[4] GARCIA Alvin, GISH Herbert: "KEYWORD SPOTTING OF ARBITRARY WORDS USING MINIMAL SPEECH RESOURCES", Proc IEEE Int. Conf. of Acoustic Speech Signal Process, ISSN : 1520-6149,Vol.2006 Vol.3 Page.949-952 (2006)

[5] Lee, S. W. etc: "Combining multiple subword representation for open vocabulary spoken document retrieval," Proc of International Conference on Acoustic , Speech and Signal Processing(ICASSP), Vol.1, pp.505-508, Mar. 2005

[6] Hiroyuki Nishi, Yoshimasa Kimura, and Nguyen Van Don: "Speech period detection using Hough transform of distance matrix images," Technical Report of IEICE, SP2008-112, pp. 197-202 (2008), in Japanese

[7] Hiroyuki Nishi, Yoshimasa Kimura, and Takeshi Ishibashi: "Word spotting framework using Hough transform of cepstrum distance matrix images," Proc. YKJCA2009, 204 (2009)

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)