STRUCTURING BASEBALL LIVE GAMES BASED ON SPEECH RECOGNITION USING TASK DEPENDENT KNOWLEDGE AND EMOTION STATE RECOGNITION

Atsushi Sako, Yasuo Ariki

Department of Computer and Systems Engineering
Kobe University, Kobe, Japan
sakoats@me.cs.scitec.kobe-u.ac.jp
ariki@kobe-u.ac.jp

ABSTRACT

It is a difficult problem to recognize baseball live speech because the speech is rather fast, noisy, emotional and disfluent due to rephrasing, repetition, mistake and grammatical deviation caused by spontaneous speaking style. To solve these problems, we propose in this paper a speech recognition method incorporating emotion state as well as the baseball game knowledge such as counting of inning, out, strike and ball. Due to this emotion state and task-dependent knowledge, the proposed method can effectively prevent speech recognition errors. This method is formalized in the framework of probability theory and implemented in the conventional speech decoding (Viterbi) algorithm. The experimental results showed that the proposed approach improved the structuring and segmentation accuracy as well as keywords accuracy.

1. INTRODUCTION

Recently a large quantity of multimedia contents are broadcast and accessed through digital TV and WWW. In order to retrieve exactly what we want to know from multimedia database, automatic extraction of meta information or structuring is required, because it is impossible to give them to multimedia database manually due to their quantity.

The purpose of this study is to automatically transcribe sports live speech, especially baseball commentary speech, in order to produce the closed caption and to structure the sports games for highlight scene retrieval. The baseball game structuring is the process of segmenting the game into a pitching sequence and giving each of them the meta information such as inning, out count, strike count and ball count. The accuracy of the baseball game structuring depends on the transcription accuracy so that sophisticated speech recognition techniques are required.

As the sports live speech, we used radio speech instead of TV speech because the radio speech has much more information about the keywords. However the radio speech is rather fast, noisy and emotional. Furthermore, it is disfluent due to rephrasing, repetition, mistake and grammatical deviation caused by spontaneous speaking style. To solve these problems, we have already proposed the adaptation techniques of language model and acoustic model, which convert a baseline model originally constructed using available speech corpus to the sports live model using sports live speech[1].

In order to further improve the speech recognition accuracy, we propose, in this paper, two new methods. The first one is an emotion state where excited announcer speech is recognized by the HMM trained by the excited training data. The second is the incorporation of a baseball task-dependent knowledge such as counting of inning, out, strike and ball. For example, in conventional speech recognition, "Ball count two and two" could occur immediately after "Ball count one and one" due to speech recognition error. Following the baseball rule, it does not occur. In the proposed speech recognition, the probability from "Strike: 1, Ball: 1" to "Strike: 2, Ball: 2" i.e. $P(2,2|1,1)$ is set to be zero. Therefore, the proposed speech recognition can prevent an incorrect recognition. These methods of emotion state and task-dependent knowledge incorporation are formalized in the framework of probability theory and implemented in the conventional speech decoding (Viterbi) algorithm.

2. KNOWLEDGE OF BASEBALL GAMES

A content of a baseball game consists of the sequence of video data and speech data. We use a commentary speech on a radio. It is spoken by an announcer watching a live baseball game. An announcer speaks based on a process of a game such as structure (inning, counts and runners) and events (strike out, base on balls, hit, out, home run and etc.) (Fig. 1). Therefore the commentary speech spoken by an announcer deeply depends on the sequence of game states. In general an announcer is excited when an event is occured. Thus, to find an event scene it is reasonable to find a scene where an announcer is excited. From the above dis-

cussion, it is easily thought that "a state" of a baseball game (more properly a commentary of a baseball game) consists of "structural state" and "emotional state". Consequently, speech recognition should be performed using game dependent knowledges presented by structural state and emotion state presented by HMM trained by excited training speech. To develop this kind of speech recognition, we propose the speech recognition method that estimates a word sequence, baseball state sequence and emotion state simultaneously.

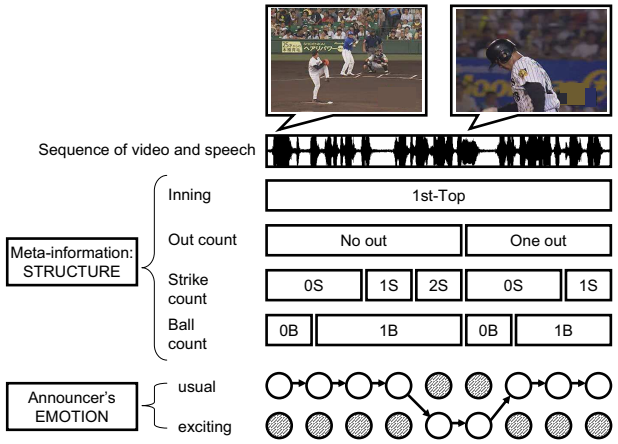

Fig. 1. *The structure of a baseball game.*

To find the correct sequence of the game states, the accuracy of speech recognition, particularly the keywords accuracy is important, because the keywords are deeply related to the structure of a baseball game. In the past, to improve the performance of speech recognition in this task, we performed the acoustic model adaptation and the language model adaptation. Herewith, the performance was quite improved, but not enough to accomplish our goal. Now we notice that the utilization of heuristic rules related to the keywords plays an important role to perform structuring of a baseball game.

3. SPEECH RECOGNITION WITH STATE ESTIMATION

In this section, we formalize the proposed speech recognition and compare it to the conventional speech recognition, through their formalization.

Let $\mathbf{O} = \{O_1, \cdots, O_T\}$ and $\mathbf{W} = \{W_1, \cdots, W_N\}$ be a sequence of D-dimensional observed feature vectors and a word sequence respectively. The general problem of speech recognition is to find the most likely word sequence $\mathbf{W}$, given the sequence of observed feature vectors $\mathbf{O}$. But our proposed method "Speech Recognition with State Estimation" is to find the most likely word sequence $\mathbf{W}$ and state sequence $\mathbf{S} = \{S_1, \cdots, S_N\}$ simultaneously, given the sequence of observed feature vectors $\mathbf{O}$ as follows:

$$(\hat{\mathbf{S}}, \hat{\mathbf{W}}) = \underset{\mathbf{S},\mathbf{W}}{\operatorname{argmax}} P(\mathbf{S}, \mathbf{W}|\mathbf{O}). \quad (1)$$

Eq.2 can be derived from Eq.1:

$$(\hat{\mathbf{S}}, \hat{\mathbf{W}}) = \underset{\mathbf{S},\mathbf{W}}{\operatorname{argmax}} P(\mathbf{S}, \mathbf{W}) P(\mathbf{O}|\mathbf{W}, \mathbf{S}), \quad (2)$$

based on Bayesian theorem and $P(\mathbf{O})$ is omitted due to independence from $\mathbf{W}$. Moreover, we can derive the following equation.

$$\begin{aligned} P(\mathbf{S}, \mathbf{W}) &= P(S_1, \cdots, S_N, W_1, \cdots, W_N) \\ &= P(S_1) P(W_1|S_1) \\ &\times \prod_{i=2}^{N} P(S_i|S_1^{i-1}, W_1^{i-1}) P(W_i|S_1^i, W_1^{i-1}). \end{aligned} \quad (3)$$

Based on the following approximation,

- A state depends only on the previous state.
- A word depends only on the previous state, present state and the previous word.

we can simplify Eq.3 as follows:

$$\begin{aligned} P(\mathbf{S}, \mathbf{W}) &= P(S_1) P(W_1|S_1) \\ &\times \prod_{i=2}^{N} P(S_i|S_{i-1}) P(W_i|W_{i-1}, S_{i-1}, S_i). \end{aligned} \quad (4)$$

Finally, the speech recognition with state estimation is formalized as:

$$\begin{aligned} (\hat{\mathbf{S}}, \hat{\mathbf{W}}) &= \underset{\mathbf{S},\mathbf{W}}{\operatorname{argmax}} P(\mathbf{O}|\mathbf{W}, \mathbf{S}) P(S_1) P(W_1|S_1) \\ &\times \prod_{i=2}^{N} P(S_i|S_{i-1}) P(W_i|W_{i-1}, S_{i-1}, S_i). \end{aligned} \quad (5)$$

Note that $P(\mathbf{O}|\mathbf{W}, \mathbf{S})$ is an acoustic model depending on the state, $P(S_i|S_{i-1})$ is a state transition probability and $P(W_i|W_{i-1}, S_{i-1}, S_i)$ is a bi-gram probability depending on the state transition. Several studies have been reported about state dependent language model in mainly spoken dialogue system[4][5][6].

Now we compare the proposed method shown in Eq.5 to the conventional method formalized as:

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmax}} P(\mathbf{O}|\mathbf{W}) \prod_{i=1}^{N} P(W_i|W_{i-1}). \quad (6)$$

Firstly, the acoustic probabilities $P(\mathbf{O}|\mathbf{W}, \mathbf{S})$ is different from the conventional method. Due to this acoustic model, it can be expected that multiple acoustical features such as usual, exciting, sad or other emotions can be represented. Especially in this paper, we expect to represent "usual" and

"exciting" speeches. Secondly, there is the state transition probability in the proposed method, not in the conventional method. Due to this probability, the proposed method can prevent a state transition not to occur under the baseball rule. Finally, the bi-gram probability depending on baseball game state $P(W_i|W_{i-1}, S_{i-1}, S_i)$ in the proposed method is similar to the bi-gram probability $P(W_i|W_{i-1})$ in the conventional method. However in the proposed method, it can be estimated between a state transition. Namely the word probability can be regarded as depending on a state transition.

Next, we describe how to learn the stochastic models in the proposed speech recognition such as the acoustic model depending on the state $P(\mathbf{O}|\mathbf{W}, \mathbf{S})$, the state transition probability $P(S_i|S_{i-1})$ and the bi-gram probability depending on the baseball game state $P(W_i|W_{i-1}, S_{i-1}, S_i)$.

4. LEARNING OF STOCHASTIC MODELS

4.1. HMM depending on a state

We use two set of HMMs to represent the state dependent HMM $P(\mathbf{O}|\mathbf{W}, \mathbf{S})$ by switching those HMMs. One represents usual emotion and the another represents exciting emotion. Each HMM is created by adaptation using speech corpus for each emotion. A speaker of adaptation corpus is the same as that of the test set. The adaptation is performed by the following method. First, emotional time sections are detected by human listening to adaptation corpus. Next, speech data and the corresponding text data at exciting time sections are separated from usual emotion time section. Next, model parameters are transformed by MLLR (Maximum Likelihood Linear Regression) using supervising text. Last, MAP (Maximum A Posteriori) estimation is performed using the transformed parameters. Here the adaptation at exciting time sections are performed after the adaptation at the usual emotion time sections because amounts of exciting time sections are small.

4.2. State transition model

In this paper, a *state* is defined as a process in a baseball game. Therefore, a state transition model presents a process flow of a baseball game. For example, in conventional speech recognition, "Ball count two and two" could occur immediately after "Ball count one and one" due to speech recognition error. This means that a state transits from "Strike: 1, Ball: 1" to "Strike: 2, Ball: 2" directly. According to the baseball rule, it should not occur. In the proposed speech recognition, the probability that a state transits from "Strike: 1, Ball: 1" to "Strike: 2, Ball: 2" i.e. $P(2, 2|1, 1)$ is set to be zero. Therefore, the proposed speech recognition can prevent an incorrect state transition.

Here, it is too difficult to learn this stochastic model from training data because a probability of state transition depends deeply on each player or baseball teams. Thus, we approximately set zero probability to a state transition not allowed by the rule of baseball. Also we set constant probability to a state transition allowed. Consequently, this state transition model represents a baseball rule network.

4.3. Bi-gram probability depending on the state transition

The bi-gram probability depending on the state transition $P(W_i|W_{i-1}, S_{i-1}, S_i)$ can be obtained through learning. However it invokes the data sparseness problems so that we focus on the following three types of commentaries on a baseball game as shown in Table 1.

Table 1. *Types of commentaries and the examples.*

	Types	Examples
(i)	Related to a state transition	"Pitch and strike." "A fly ball was caught. Out."
(ii)	Explain a game state	"Count is two and two." "Two out and two on."
(iii)	No relation with a game state	"I'll compare this game to the previous one." "It is cloudy with a chance of rain."

Type (i) is related to a state transition. If an utterance of this type occurs, it leads to the transition of a state. For example, utterance of "Pitch and strike." leads to the state transition that increases a strike count.

Type (ii) is for explaining a game state. For example, if utterance of "Count is two and two." is spoken, it indicates high probability that a state is in (Strike: 2, Ball: 2) now. Our experience shows that utterances of this type can correct a wrong state transition.

Type (iii) is not related to a state. An utterance in this type doesn't lead to state transition nor explain a state. So, it can be thought that a sequence of words $\mathbf{W}$ doesn't depend on a sequence of states $\mathbf{S}$. Therefore, in this type, $P(W_i|W_{i-1}, S_{i-1}, S_i)$ can be reduced to $P(W_i|W_{i-1})$. This equation is same as the bi-gram probability in the conventional speech recognition.

For type (i) and (ii) commentaries, we manually marked the game state to the training data, and computed the bi-gram probability statistically from this training data. In the type (iii), we used a bi-gram same as the conventional speech recognition.

5. EXPERIMENTS

5.1. Experimental conditions

The proposed speech recognition with state estimation described in Section 3 enables incorporating baseball dependent knowledge into speech recognition and also enables visualizing a sequence of baseball game states and time sec-

tions of speaker's emotion. We carried out experiments to prove the effectiveness of the proposed speech recognition.

The experimental conditions are summarized in Table 2. In order to improve the speech recognition accuracy for a commentary on a baseball, we employed the acoustic and language model adaptation. In the acoustic model, the training data for a baseline consisted of about 200,000 Japanese sentences (200 hours) spoken by 200 males in CSJ (Corpus of Spontaneous Japanese)[2]. The adaptation was performed based on the method derived in Section 4.1. In the language model, the training data for a baseline consisted of 570,000 morphemes collected from web pages about baseball. This was named "Web text corpus". We also made "Dictated text corpus" that was manually transcribed from commentary speech on a baseball game. Then, we merged these two corpora into one corpus named "Merged text corpus". Finally, we further merged "Dictated text corpus" and "Merged text corpus" with weighting under condition of minimum perplexity. We selected the keywords deeply related to the structure of a baseball game as shown in Table 3. The number of baseball states was 72 (3 strikes, 4 balls, 3 outs and 2 emotions). Experiments were carried out under these conditions, using decoder based on ML back-off[3].

Table 2. *Experimental conditions*

Sampling rate/Quantization	16 kHz / 16 bit
Feature vector	26 - order MFCC
Window	Hamming
Frame size/shift	20/10ms
# of phoneme categories	244 syllable
# of mixtures	32
# of states (Vowel)	5 states and 3 loops (Left to Right)
# of states (Consonant+Vowel)	7 states and 5 loops (Left to Right)

Table 3. *Keywords*

Strike, Ball, Four ball (Base on Balls in English)
Missed swing, Strike out, Foul, Out

5.2. Experimental result

Table 4 shows the experimental results. "Baseline" indicates a result by conventional speech recognition using the acoustic model and language model adaptated. "Proposed method" is a result by our method incorporating baseball task dependent knowledge into speech recognition. Note that "correct rate of structuring" is a percentage of the correctly recognized structure (inning, out count, strike count, ball count) to the total number of structure and "correct rate of detecting excited scenes" is a percentage of the correctly recognized time sections in excited speech. We can improve the keyword accuracy by 1.1%. This is mainly because the proposed method prevented incorrect words based on baseball dependent knowledge. For example, there is an utterance such as "... foul ball, and strikeout in next pitch". In conventional speech recognition, it misrecognized "four ball (Base on Balls in English)" instead of "foul ball" due to very similar pronunciation. But the strikeout should not occur immediately after four ball because four ball sets the strike count to zero. The strikeout can occur under the condition of strike count two. The proposed speech recognition can estimate a sequence of game states and correctly recognize "foul ball, and strikeout". We confirmed 71.4 % of the correct rate of structuring and 75.0% of the correct rate of detecting excited scenes.

Table 4. *Experimental results*

	Baseline	Proposed method
Keyword accuracy	66.8 %	67.9 %
Correct rate of structuring	-	71.4%
Correct rate of detecting excited scenes	-	75.0%

6. SUMMARY

In this paper, we described how to structure the commentary speech by incorporating baseball task dependent knowledge and proposed the speech recognition method with state estimation. It can be thought that the proposed speech recognition is a sort of the information integration. The experimental result showed that the proposed method improved the keyword accuracy, and herewith achieved the 71.4% of the correct rate of structuring and 75.0% of the correct rate of detecting excited scenes.

7. REFERENCES

[1] Y. Ariki, T. Shigemori, T. Kaneko, J. Ogata and M. Fujimoto: "Live Speech Recognition in Sports Games by Adaptation of Acoustic Model and Language Model", Eurospeech2003, pp.1453-1456, 2003-09.

[2] S. Furui, K. Maekawa, H. Isahara: "Spontaneous Speech: Corpus and Processing Technology", The Corpus of Spontaneous Japanese, pp.1-6, 2002-2.

[3] J. Ogata, Y. Ariki: "An Efficient Lexical Tree Search for Large Vocabulary Continuous Speech Recognition", Proc. of the Sixth Int'l Conf. on Spoken Language Processing(ICSLP'00), Vol.II, pp.967-970 (Oct. 2000).

[4] C. Popovic and P. Baggia: "Specialized language models using dialogue predictions", in ICASSP 1997, pp 423-426, 1997.

[5] F. Wessel and A. Baader: "Robust dialogue-state dependent language modeling using leaving-one-out", in ICASSP 1997, pp741-744, 1997.

[6] Wei Xu and Alex Rudnicki: "Language modeling for dialogue systems", in ICSLP 2000, pp 118-121, 2000.