

Live Speech Recognition in Sports Games by Adaptation of Acoustic Model and Language Model

Yasuo Arika, Takeru Shigemori, Tsuyoshi Kaneko, Jun Ogata, Masakiyo Fujimoto

Department of Electronics and Informatics
Ryukoku University, Japan
ariki@rins.ryukoku.ac.jp

Abstract

This paper proposes a method to automatically extract keywords from baseball radio speech through LVCSR for highlight scene retrieval. For robust recognition, we employed acoustic and language model adaptation. In acoustic model adaptation, supervised and unsupervised adaptations were carried out using MLLR+MAP. By this two level adaptation, word accuracy was improved by 28%. In language model adaptation, language model fusion and pronunciation modification were carried out. This adaptation showed 13% improvement at word accuracy. Finally, by integrating both adaptations, 38% improvement was achieved at word accuracy level and 28% improvement at keyword accuracy level.

1. Introduction

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

Multimedia contents can be classified into two groups; one is text-oriented content such as news, documents and drama which are produced according to the well-organized story text. The other is event-oriented content such as sports live video produced through player's action and announcer's live speech. The former contents can be indexed using the text information. However, the latter contents require a quick indexing through automatic speech recognition.

The purpose of this study is to automatically transcribe sports live speech in order to produce the closed caption, to structure the sports games and to extract keywords for highlight scene retrieval. The accuracy of the sports game structuring and the highlight scene retrieval depend on the transcription accuracy so that sophisticated speech recognition is required.

As the sports live speech, we used radio speech in stead of TV speech because the radio speech has much more information about the keywords. However the radio speech is rather fast and noisy. Furthermore, it is disfluent due to rephrasing, repetition, mistake and grammatical deviation caused by spontaneous speaking style. To solve these problems, we propose in this paper adaptation techniques for language model and acoustic model, which converts a baseline model originally constructed using available speech corpus to the sports live model using sports live speech.

We describe the acoustic model adaptation, language model adaptation and keyword extraction method from speech data in the following sections.

2. Live speech features and adaptation

Table1 shows a comparison of speaking styles among read speech such as in news and document, lecture speech and live speech. As can be seen, the live speech is noisy, emotional and unclear due to its high speaking rate compared to the other speaking styles. In addition, the live speech is disfluent due to repeat, mistake and grammatical deviation.

In our case, the radio speech was recorded in a relatively quiet booth so that the environmental noise is not so strong. From the above described features of live speech, we constructed the acoustic model by using lecture corpus (CSJ: Corpus of Spontaneous Japanese) including 200 male speakers[1] because there was no baseball speech corpus in the world yet. Then the constructed acoustic model was used as a baseline in speech recognition. The baseline acoustic model is converted to live speech model by adaptation techniques using the live speech data.

The language mode was constructed using baseball text corpus that was originally collected through WWW because there was no baseball text corpus in the world yet. Then the constructed language model was used as a baseline in speech recognition. The baseline language model is converted to live language model by adaptation techniques using live speech transcription. Hereafter, we describe the employed techniques for the acoustic model adaptation and language model adaptation.

Table 1: Comparison of speaking styles

	Speaking rate	Noise	Emotion
Read speech	7.26 (mora/sec)	Quiet	Weak
Lecture speech	7.31 (mora/sec)	Middle	Middle
Live speech	8.51 (mora/sec)	Noisy	Strong

3. Acoustic model adaptation

3.1. Supervised adaptation

Fig.1 shows the acoustic model adaptation process. The baseline acoustic model (HMM:Hidden Markov Model) was constructed using lecture corpus so that the model is not suitable for the live speech recognition. In order to absorb the difference in both of the speaking style and speakers, the baseline HMM is converted to the adapted HMM by supervised adaptation which utilizes adaptation speech and manually transcribed text data.

The adaptation speech was collected from one baseball game (70 minutes) and manually transcribed. The

adaptation method is MAP (maximum a posteriori probability) adaptation[2] after MLLR (Multiple linear regression) adaptation[3]. The MLLR adapts the baseline HMM quickly to the target speaking style owing to Affine transformation and the MAP adapts it precisely to the target speaking style based on a posteriori probability.

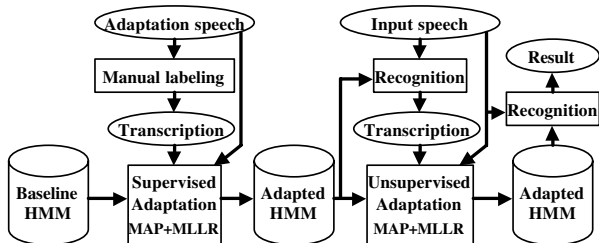


Figure 1: Acoustic model adaptation process

3.2. Unsupervised adaptation

The adapted HMM is suitable for the baseball live speech, but speech data for evaluation is slightly different from the adaptation speech in speaking style, speaker characteristics and environment noises. In order to absorb this difference, the adapted HMM is further adapted to the speech data for evaluation by an unsupervised technique utilizing input (evaluation) speech and automatically recognized transcription.

The adaptation method is also MAP after MLLR here. The difference between supervised and unsupervised adaptation lies in the difference of used transcription; manual transcription or automatically recognized transcription. It is clear that the accuracy of supervised adaptation is superior to the unsupervised adaptation.

4. Language model adaptation

4.1. Text corpus

Baseball text corpus was originally constructed by collecting baseball text through WWW because there has been no baseball corpus in the world yet. We call this baseball corpus collected from WWW as *web text corpus* in this paper. The size of the web text corpus is 576,936 words. The collected text are parsed into words through a morphological analysis by ChaSen[4] and bigram / trigram language models as well as the dictionary are constructed using CMU-Cambridge Toolkit[5].

The baseline language model is constructed using this web text corpus. The web text corpus is a collection of text in a written style so that the baseline language model is not suitable for live (spontaneous) speech recognition. From this viewpoint, the baseline language model is converted to live language model by adaptation techniques using the live speech transcription.

We call the corpus used for this adaptation as *adaptation text corpus*. The size of the adaptation text corpus is 10,865 words. The transcription of the adaptation text corpus is produced manually for one baseball game (70 minutes). The transcription is parsed into words through a morphological analysis by ChaSen. Then the bigram and trigram language models are constructed using CMU-Cambridge Toolkit as well as the dictionary.

The *joint text corpus* is available by merging two corpora, web text corpus and adaptation text corpus. This joint text cor-

pus also can be used for language mode adaptation.

4.2. Language model fusion

Two language models constructed using the web text corpus and the adaptation text corpus respectively can be fused into an adapted language model by a method described in the paper[6]. However, in our case, two kinds of corpora are available so that the following three types of language model fusions are feasible.

- (1) A language model fused by two language models constructed using Web text corpus and adaptation text corpus respectively.
- (2) A language model fused by two language models constructed using join text corpus and adaptation text corpus respectively.
- (3) A language model constructed using joint text corpus.

We verify experimentally which language model is best for language mode adaptation in sec6.2.

4.3. Concept classes

In the sports live speech, player names and commentator names are usually observed. However, their frequency is not so large to reflect them into the language model. To solve this problem, two classes are employed; one is PLAYER class for a collection of player names and the other is COMMENTATOR class for a collection of commentator names.

The language model (bigram or trigram) is constructed using the class names instead of individual player or commentator name. Namely, in the web text corpus and adaptation text corpus, the player names and commentator names are converted to the corresponding class names PLAYER and COMMENTATOR. In speech recognition, the bigram language probability is used for transition from a word to the class names and the acoustic model probability is computed for all the player names within the class name.

4.4. Modification of pronunciation dictionary

As mentioned earlier, the live speech is fast so that the pronunciation is deviated from the normal one and this causes the speech recognition errors. To solve this problem, the pronunciations of some words in the dictionary are modified manually to absorb the deviation. For example, the pronunciation /bo: ru ka u N to/ for the word ball-count is modified to /bo: ru ka u N/ because, in fast speech, the last vowels of the words are sometimes omitted.

5. Speech recognition

5.1. Speech recognition system

We employed a 2-pass decoder as a LVCSR (large vocabulary continuous speech recognition) system as shown in Fig.2[7]. At the 1st-pass, we adopted a lexical tree search using a bigram language model for constructing the word graph. A search method we employed is called "best-word back-off connection" which has been already proposed[8]. This method links the word with the best partial score at each frame to the back-off connection so that it can reduce about half of the processing time without increasing any errors. At the 2nd-pass, the best sentence (word sequence) is searched in the word graph using a trigram language model.

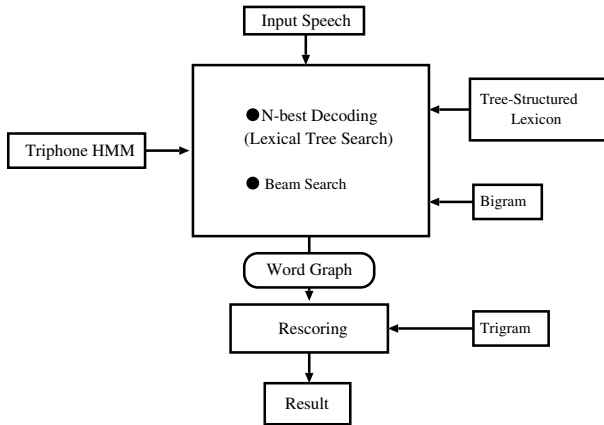


Figure 2: Speech recognition system

5.2. Keyword extraction

Since one of the purpose of this study is to extract highlight scenes, the keywords related to highlights have to be prepared in advance. The highlight scene is defined as the scene strongly concerning with the score. From this viewpoint, we prepared the keywords shown in Table2.

Keywords are sometimes observed at irrelevant time before or after the highlight scenes. For example, the announcer remembers the home run and refers to it a little later after it has finished. The difference between the keywords at relevant time and irrelevant time can be observed in the difference of the emotion, especially in the power of the speech.

From this viewpoint, we extract keywords with their time sections and then the power of the keywords is computed within the time section. If the power is bigger than some threshold, then the keyword is confirmed as a true keyword. Otherwise they are rejected as false keywords.

Table 2: Keyword list

Home run, Two-base hit, Three-base hit, Bases loaded, Grand slam, Timely hit, Home steal, Insurance run, The points scored first, Bases-loaded walk

6. Experimental Results

6.1. Experimental condition

We carried out acoustic model and language model adaptation using the adaptation data, and also carried out speech recognition and keyword extraction experiments for the evaluation data both shown in Table3. In the table, the number / number shows the month and day when the game was played.

The baselines for the acoustic model and language model were constructed using CSJ (lecture) speech corpus and web text corpus described in Sec.3 and Sec.4 respectively. Table4 shows the condition for acoustic analysis (AA) and HMM.

6.2. Best method for language model adaptation

Table5 shows the comparison of three types of language model adaptations by speech recognition at word accuracy level for

Table 3: Speech data for adaptation and evaluation

Set	Adaptation data	Evaluation data
1	9/8	9/24
2	8/29	10/6

Table 4: Condition for acoustic analysis and HMM

	Sampling frequency	16KHz
	Feature parameters	MFCC(39 dim)
A	Frame length	20ms
A	Frame shift	10ms
	Window type	Hamming
H	Acoustic unit	244 Syllables
M	Mixture Num	32
M	Vowel	5 states with 3 loops
M	Consonant+Vowel	7 states with 5 loops

two game sets. In the table, the method1, method2 and method3 correspond to the methods described in sec4.2.

From the table, the method2, which fuses two language models constructed using joint text corpus and adaptation text corpus, showed the best word accuracy. This can be explained that the adapted language model is rather shifted to the adaptation data because the joint text corpus itself includes the adaptation text corpus. In the following experiment, this type of language model adaptation is employed.

Table 5: Comparison of language models by word accuracy(%)

	TestSet	
	1	2
Method1	48.7	33.9
Method2	51.9	38.4
Method3	49.4	36.1

6.3. Result of language model adaptation

Table6 shows the result of speech recognition for two game sets before and after the language model adaptation. In the table, baseline shows the speech recognition result using the baseline language model constructed by the web text corpus while using the baseline acoustic model constructed by CSJ (lecture) speech corpus. LA shows the speech recognition result after the language model adaptation using the concerning adaptation data.

P.P. shows test set perplexity which shows the language complexity. OOV shows the percentage of the words out of vocabulary. Corr and Acc show the word correct rate and word accuracy which show the speech recognition performance. KW-Corr and KW-Acc show the keyword correct rate and keyword accuracy before the power discrimination described in Sec.5.2.

From the table, it can be seen that by the language model adaptation the P.P. significantly decreased to about 30%, and Corr and Acc were significantly improved by about 13%. The keyword accuracy was improved but the keyword correct rate was not improved. This indicates that the language model con-

tributes to reduce the falsely accepted keywords, but does not contribute to improve the missed keywords.

Table 6: Result of language model adaptation

TestSet	1		2	
	Baseline	LA	Baseline	LA
PP	258.4	75.8	248.7	69.2
OOV	6.9	4.2	11.12	7.0
Corr	43.8	58.3	38.3	49.2
Acc	35.1	51.9	27.3	38.4
KW-Corr	82.0	81.6	80.0	76.7
KW-Acc	53.8	78.9	74.3	76.7

6.4. Result of acoustic model adaptation

Table7 shows the speech recognition result before and after the acoustic model adaptation while keeping a baseline language model. In the table, the baseline shows the result before the acoustic model adaptation. On the other hand, AA1 and AA2 show the result after the acoustic model adaptation by supervised and unsupervised methods respectively.

From the table, it can be seen that the word accuracy was improved by almost 30%, and AA2 shows 3% improvement compared with AA1. This indicates that the supervised and unsupervised acoustic model adaptation are both effective. The keyword accuracy was improved by almost 22%. These improvements are attributed to the effectiveness of the speaker and environmental noise adaptation. In the following experiment, the acoustic model is employed after the supervised and unsupervised adaptation.

Table 7: Result of acoustic model adaptation(%)

TestSet	1			2		
	Baseline	AA1	AA2	Baseline	AA1	AA2
Corr	43.8	70.0	70.9	38.3	64.0	66.0
Acc	35.1	62.7	64.0	27.3	52.6	55.6
KW-Corr	82.0	92.0	94.1	80.0	97.1	97.1
KW-Acc	53.8	84.0	84.3	74.3	88.2	88.2

6.5. Integrated result of model adaptation

Table8 shows the integrated speech recognition result before and after the adaptation of both acoustic model and language model. In the table, the baseline and LA-AA show the result before and after the adaptation. From the table, it can be seen that the word accuracy was improved by almost 38% and the keyword accuracy was improved by almost 28%.

Table 8: Integrated result of model adaptation(%)

TestSet	1		2	
	Baseline	LA-AA	Baseline	LA-AA
Corr	43.8	78.6	38.3	72.8
Acc	35.1	74.6	27.3	63.8
KW-Corr	82.0	93.9	80.0	97.1
KW-Acc	53.8	87.8	74.3	97.1

6.6. Result of keyword extraction

Table9 shows the keyword extraction rate after the power discrimination described in Sec.5.2. By the power discrimination, the false keywords are significantly reduced and the true keywords are successfully extracted. However, there are still false extraction and missing keywords. They can be explained partly because the acoustic model is still weak and partly because the power discrimination method has its limitation that the keywords occurring at the irrelevant time with strong power are sometimes extracted as true keywords.

Table 9: Result of keyword extraction

Set	Extraction (Correct #/True #)	False #
1	2/2	2
2	2/4	0

7. Conclusion

In this paper, we proposed the method to automatically extract keywords from baseball radio speech through LVCSR for highlight scene retrieval. For robust recognition, we employed acoustic and language model adaptation. In acoustic model adaptation, supervised and unsupervised adaptations were carried out using MLLR+MAP. By this two level adaptation, word accuracy was improved by 30%.

In language model adaptation, language model fusion, concept class generation and pronunciation modification were carried out. This adaptation showed 13% improvement at word accuracy. Finally, by integrating both adaptations, 38% improvement was achieved at word accuracy level and 28% improvement at keyword accuracy level. In future, we are going to study further adaptation techniques and keyword extraction methods.

8. References

- [1] K.Maekawa, H.Koiso, S.Furui and H.Isahara: "Spontaneous speech corpus of Japanese", Proc. of LREC2000, pp.947-952, Athens, 2000.
- [2] J.L.Gauvain and C.H. Lee: "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains", IEEE Trans. on Speech and Audio Processing, Vol.2, no.2, pp.291-298, 1994.
- [3] C.L.Leggerter and P.C.Woodland: "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models", Computer Speech and Language, Vol.9, pp.171-185, 1995.
- [4] Morphological Analyzer ChaSen
<http://chasen.aist-nara.ac.jp/index.html>
- [5] P.R.Clarkson, R.Rosenfeld: "The CMU-Cambridge Statistical Language Modeling Toolkit v2",
<http://svr-www.eng.cam.ac.uk/prc14/toolkit.html>
- [6] R.Nishimura, K.Komatsu, Y.Kuroda, K.Nagatomo, A.Lee, H.Saruwatari, K.Shikano, "Automatic N-gram Language Model Creation from Web Resources", Proceedings of 7th European Conference on Speech Communication and Technology, pp.2127-2130, September 2001
- [7] S.Ortmanns, H.Ney and X.Aubert, "A Word Graph Algorithm for Large Vocabulary Continuous Speech Recognition", Computer Speech and Language, Vol.11, No.1, pp.43-72, 1997.
- [8] J.Ogata, Y.Ariki, "An Efficient Lexical Tree Search for Large Vocabulary Continuous Speech Recognition", ICSLP'00, Vol.II, pp.967-970, 2000.