Speaker Labeling Using Closed-Captioning

Keita Yamamuro
Hosei University
Tokyo, Japan
Email: keita.yamamuro.xn@stu.hosei.ac.jp

Katunobu Itou
Hosei University
Tokyo, Japan
Email: it@fw.ipsj.or.jp

***Abstract*—There has recently been much research on annotation systems for television broadcasting because of interest in retrieving highlights from television programs. However, most of the methods developed have specialized in only one genre. Therefore, in this study we targeted three genres drama, animation, and variety and developed a system of annotating indexical information through metadata obtained from television captions. Specifically, the information from the captions is used to create a phoneme HMM that is then used for speaker identification. The proposed system selects the most appropriate phonemic model from several candidate models based on the Bayesian information criterion (BIC) of likelihood and data. Characters in 70 television programs were identified with a recognition accuracy of 39.6%. Television captioning can already identify about 50.0% of the characters in a show, and when we combined captioning with the proposed system, 70.0-80.0% of the utterances in one program were correctly identified.**

***Keywords*-Speaker Identification; Model Selection; HMM; Television Broadcasting; Speaker Diarization.**

I. INTRODUCTION

Recently, an incredible amount of video content is being broadcast by multi-channel services, and images that viewers demand cannot immediately be obtained. There are a few different kinds of server-type broadcasting services. Most use home servers that can retrieve and store television programs and metadata [9]. Such services access the metadata and retrieve highlights. Scene information in the metadata is crucial for retrieval and editing, but it is extremely time consuming to manually extract it. In response to this problem, several systems to efficiently process metadata have been proposed. These automated systems combine image-recognition, speech-recognition, and natural-language processing technologies [1], [2], [3], [4]. In this study, we propose a system of annotating indexical information through metadata obtained from television captions.

The remainder of this paper will proceed as follows. First, we introduce the background research in Section Ⅱ. Subsequently, we detail methods of speaker identification by model selection in Section Ⅲ. We present an experimental setup and results in Section Ⅳ. We present conclusions and future work remarks in Section V.

II. BACKGROUND RESEARCH

Many methods of annotating information from television broadcasts have previously been proposed. These methods use various technologies, including image recognition, speech recognition, and natural-language processing, to retrieve television programs, edits content, and enhances speech.

For example, information retrieval interface that uses image processing [5]. The video has been annotated by image recognition. And that information will be searched by keyword.

Methods of analyzing speech in news broadcasts have been particularly researched [6], [7] because news programs contain quite varied information: speaker information is often crucial when trying to pinpoint the relevant information in a large amount of content. Annotation of scene information from sports programs has also been extensively researched [8], and program highlights are selected and edited based on the relevant scene information.

Much of the research described above is targeted at either news or sports programs. Speech from news programs does not contain a great deal of noise, as it is clearly uttered, and speech in sports programs can be easily processed with specific key words like “ goal” or “ shot.” In other words, analyzing these types of shows is not terribly difficult. However, many other types of television programs are more complex and require painstaking annotation. In this study, we targeted drama, animation, and variety shows and developed a method that can identify speaker information within them.

III. SPEAKER IDENTIFICATION

The purpose of this research was to come up with a way to annotate indexical information in television programs. The proposed method can do this by using not only a program’s audio track but also its closed television captions. The annotation is performed by processing metadata obtained from television captions and speaker identification. In general, 50.0% of the utterances in a typical television program are annotated by television captions because this is the percentage that contains implicit speaker information. The proposed method identifies who utters the remaining 50.0%. Title information is used to determine the speaker’s identification. The proposed method constructs HMMs of all phonemes on the basis of the content of the utterance in the caption information. An HMM identifies the speaker in each phoneme of one utterance if the appropriate model has

 ISBN: 978-1-61208-157-1

been selected. Obviously, the identification rate is improved if the best model is selected. We perform this selection using decentralization, likelihood, and the amount of data in the models.

A. Metadata

Metadata contain the title, category and genre of the television program being broadcast. For example, they contain plot outlines, the names of performers and producers, and broadcasting times. Metadata has the advantage of being easily searchable, so there has been much research on the different types of metadata, e.g., that of a soccer program [9]. In this example, the metadata includes the content of the game based on the classification of actual keywords (e.g., "goal" and "kickoff"). There has also been research on classification according to topics in news programming [2]. Such topics are classified depending on the image, speech, and natural-language processing metadata. This research has made it possible for users to retrieve the scenes they want from the metadata. The purpose of that study was to analyze information from utterances in television programs: for example, "who is speaking?", "when are they speaking?", and "what are they saying?"

B. Television captions

The Ministry of Internal Affairs and Communications is working to spread the use of captioning in television broadcasting throughout Japan [10]. The present percentage of broadcast captioning is 40.1% of the total broadcasting time. The aim is to boost this percentage up to 69.9%, thus providing wider access to the information it contains.

This captioning information includes the time and content of the captioning displays as well as the kinds and colors of fonts used [11]. Television captions for main characters are in color. This means that some speakers can be identified by the font information. However, at present this type of identification works only about 40-60% of the time because often there is television captioning with no speaker information and no utterances. Although its use is limited, speaker information can be effective at extracting scenes and identifying people. In this study, we classified speech in television broadcasting by using information from the television captioning.

Although the speech of all characters contains speaker information, this is only applies to some speeches. We developed a method to identify speakers using these data.

In the proposed method, information from television captioning is used to create a model for speaker identification. Captions have information about the beginning and end of a speech, and speech that identifies the speaker is extracted from this information. When a speaker is successfully identified, the extracted speech is used for training data, and if it is uncertain who the speaker is, it is used as test data. Several speakers can also be identified by the color of the font and the content of the television captioning. Speech by main characters is in colored font, and because a speaker's speech is all in the same color, all of his/her speech can be identified. Information on the speaker might be included in the content of television captioning where the name of the speaker is in parentheses. Because other speech by the speaker cannot be identified, this speech is used as training data.

The content of television captioning is useful to construct a model that identifies the speaker. Because the content of the speech is understood, the speech can be analyzed phoneme by phoneme.

C. Phoneme alignment

The proposed method uses phoneme alignment to identify speech from phonemes. The speaker model in the present study was constructed with this phoneme information. We used Julian, which is speech recognition decoder software, for the alignment. Julian is open-source and high-performance software for large vocabulary continuous speech recognition decoders used by speech-related researchers and developers [12]. Julian aligns phoneme units to obtain speech-recognition results. The frames of the phoneme boundaries and the average acoustic scores per frame are calculated. The phoneme alignment in this study gives the verbal information to be identified beforehand according to the television captions. The analytical precision of the phoneme alignment is improved by giving prior information. The results of this analysis were used in this study.

D. Speaker identification model

In this study, we used a hidden Markov model (HMM) as the speaker model. HMM is a probabilistic model. It is used to represent the voice feature amount. Speaker identification is determined by the likelihood of the HMM and the input speech. The speech data from television broadcasting was used to train the HMM. The speaker model of each speaker was trained by using phoneme sections analyzed by alignment. The model in three states handled the HMM of phoneme units. There were 35 kinds of phonemes. The procedure to train the model was as follows.

First, voice activity detection was executed by using television captioning to determine the different sections of phoneme alignment. This detection uses information from the beginning and end of a television caption.

Next, phoneme alignment was used to extract phoneme information to construct the phonemic model. The preprocessing during the phoneme alignment converted the content of television captions into phonemes through morphological analysis. Sections of voice corresponding to phonemes were distinguished by phoneme alignment based on these sounds. Information on each phoneme was used to extract the mel-frequency cepstral coefficient (MFCC) features for either training or test data. MFCCs from known speakers were

 ISBN: 978-1-61208-157-1

used to train the HMM while those for uncertain speakers were used as test data.

Finally, speaker features were trained to the speaker model with the training data obtained from the television captions. All characters' names appear at some point in the captions, and these names plus utterance information are used for the initial data training. The amount of training data differs depending on the program, which means that there is always the possibility of insufficient data. We successfully identified various speakers by using this model.

However, alignment could not be used to analyze all speech. For instance, sometimes noise overlapped with voices, or there was a discrepancy between the captions and the speech. Because the alignment failed with this analysis, the phoneme unit model could not be used to accurately identify all speakers. In cases where the alignment failed, we were able to identify speech with a Gaussian mixture model (GMM) of one state and 32 mixtures. This model was trained with the average features of all phonemes.

We used the hidden Markov model toolkit (HTK) for feature extraction and for the HMM and GMM training [13]. HTK, a software toolkit for building speech recognition systems using continuous density hidden Markov models, is what the proposed method uses to analyze speech data. We also used it to construct speaker models. The constructed model identifies the speaker by HTK. The identified speech data is television captions that cannot specify the speaker. This data makes up 40.0-60.0% of all utterances. The proposed method identifies this data with HTK and then labels it.

E. Mel-frequency cepstral coefficient (MFCC)

There are some kinds of speech feature. In the present study, the MFCC is used as a feature for the speaker identification. The MFCC is analyzed by a filter bank on the Mel frequency axis. And, spectral analysis of the output from multiple filters arranged along a frequency axis is carried out. The MFCC is provided by discrete cosine transform (DCT) of the power in each band obtained in the result. The human ear is fine-tuned for hearing low-frequency sounds but not high-frequency sounds. The feature of MFCC is the same as the feature. Even if the same phoneme is analyzed, MFCC shows the feature that varies from person to person.

F. Effective model selection

Although the speaker model could identify speakers from the test data MFCCs, the amount of phoneme data used for training varied depending on the phoneme. This means there might not be enough data to learn all the phonemes in the trained speaker model.

Recognition model selection [14] and noise model selection [15] have been proposed as model selection methods. In this study, we used BIC, which is an information criterion for the performance of a probabilistic model corresponding to the amount of training data. There are other information criteria similar to BIC that have previously been proposed [16]. In many cases, the performance of a model is measured with these standards and then the most appropriate model is selected. In this study, we used 35 phoneme units as candidates and the top five were used for speaker identification. Even if there was not much training data, the proposed system could identify the speaker according to the model that was selected.

G. Assessment of identification results

We used the proposed method to select five models for the speech identification. However, the identification results were not all the same: different identification results were output by all models in cases where the system could not specify who the speaker was. In this study, we required just one identification result, which we obtained by comparing the likelihoods of all speakers being identified.

First, the identification results of one phoneme were output corresponding to the number of candidates. All likelihoods were output at the same time. Similar processing was executed for one speech. Next, the output likelihoods of all candidates were totaled. The candidate with the highest total was assumed to be a correct identification result.

IV. EXPERIMENTAL SETUP

A. Description of experiment

We performed experiments on the speaker identification system to evaluate the accuracy of the annotation used for scene retrieval. We prepared assessment data and performed speaker identification with the constructed model. In this study, we annotated information from recorded television broadcasts. We did not have much data to begin with, so we gradually added more as we went through other broadcasts. The proposed method was evaluated by comparing its performance with that of a model already known to be effective. We deemed models effective if they could identify all the speech in a particular television program. The evaluation data were extracted from drama, animation, and variety programs broadcast in Japan. In total, we used 70 programs: 20 dramas, 30 animations, and 20 varieties. Shows that were 30 minutes long averaged about 400-600 sentences, and shows that were 60 minutes long averaged about 900-1100. About 40-60 % of the speech was identified by television captioning and the remainder by the proposed method. We stipulated that speech used for training must consist of at least one sentence. Table I summarizes the analytical conditions for the speech. Speaker identification was performed by the selected model and the method using all the phonemic models was evaluated at the same time for the sake of comparison. BIC identified the trained model as the best. If alignment failed, speech was identified by the GMM described earlier.

We used different episodes from the same television programs to obtain additional character data. For example, if characters from a drama in a second episode appeared

 ISBN: 978-1-61208-157-1

Table I
ANALYSIS CONDITIONS

Number of television programs	70
Number of identified people (Average from one television program)	10
Time television program lasted	30 min, 1 or 2 hr
Sampling frequency	16 kHz
Quantization bit rate	16 bits
Frame period	10 ms
Frame length	25 ms
Feature amount	MFCC (1-12) logarithm power (1) +Δ (total 26 dimensions)

in the first episode, data from the first one were added to the second one as training data. The character data therefore increases if characters appear several times. This technique allowed us to double our training data.

B. Experimental results

Figure 1 shows the experimental results obtained by adding data. Additional data improved the average identification accuracy by 2.84% for dramas, 3.41% for animation, and 1.16% for variety. This clearly demonstrates that using speech data from another episode of a program as training data improves the identification accuracy.

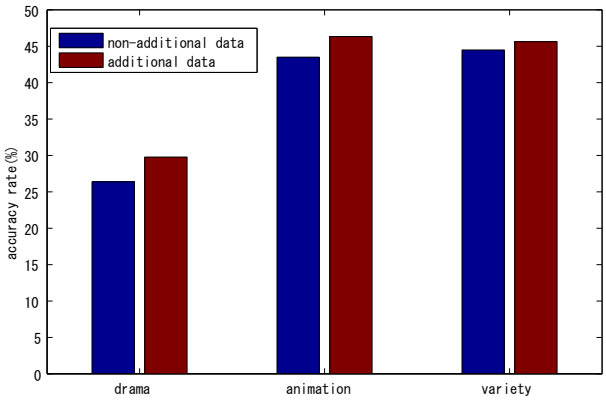

Figure 1. Results from additional data

Next, Fig. 2 shows the experimental results for model selection. Model selection improved the average identification accuracy by 11.75% for dramas, 15.15% for animation, and 15.85% for variety. 44.48% of characters were accurately identified in variety shows, which was the strongest result in this experiment. However, other studies have shown that 80.0-90.0% of speakers in news programs can be identified [17]. It is more difficult to identify speakers in dramas, animation, and variety shows than in news programs because in news programs there is less background music (BGM), fewer speakers, and clearer speech. In our experiment, we could not identify speech very well when there was an overlap of BGM and sound effects (SEs). We need to adapt the proposed method so that it can deal with such problems.

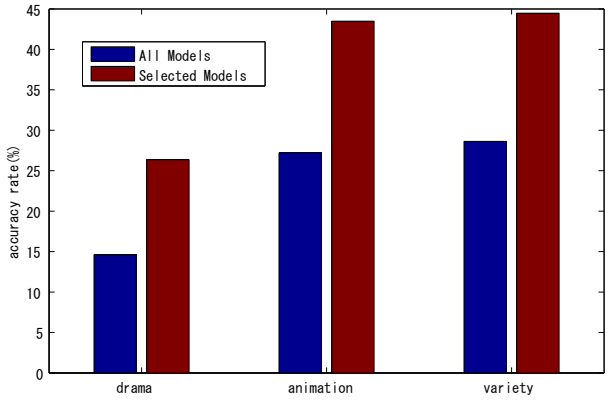

Figure 2. Results from model selection

C. Discussion

The identification accuracy improved with additional data. The identification models had different performances before and after the data were added. For example, some of them could not identify speaker features at the beginning, when there was not sufficient data, but then dramatically improved when the amount of data was increased. Take the case of a model that could accurately identify vowels such as /a/ and /i/. In this case, the model was well selected because these vowels can be used to identify speaker features when there is not much data. However, when the amount of data increased, the identification model had more freedom to use data other than vowels (e.g., consonants like /k/ and /t/). The additional data gave the model more choices, which led to improved identification accuracy.

Accurate model selection also increased the percentage of speakers who were identified. As stated previously, the proposed method selects the most effective model from among several candidate models. However, sometimes selection was difficult because each phonemic model's performance changed depending on the speaker if there was not much training data. It is therefore important that the model used for speaker identification considers all the phonemic models.

The proposed method did not deal well with overlapping BGM and SE, probably because speaker features might have been cancelled due to BGM. The identification accuracy might improve if speech is enhanced or noise is rejected.

The proposed method also selects a second-best model, so if the first model fails to identify a speech, this second one might be able to. Comparisons between the first and second models showed that there was usually not much difference in terms of identification likelihood. Therefore, identification results were requested from both models when there was

 ISBN: 978-1-61208-157-1

little difference. As a result, the recognition accuracy in all television programs was improved, in some cases by as much as about 10.0%. This demonstrates the importance of considering the difference in likelihoods.

V. Conclusions and future work

The speakers in 70 television programs were identified through television captioning information and model selection in the present study. The 44.48% correct identification rate for variety shows was the strongest result. This identification improved by 17.10% due to model selection. Overall, the proposed method was effective, but it was deficient when compared with the conventional one in terms of identifying speech in news programs.

We need to re-think how the model selection is performed. In this study, we selected models by using BIC, but because the best model changes depending on the speaker, the speaker identification should really use all the models. Ideally, the model selection should analyzes the model performances individually for each speaker and then attach the appropriate weight to the result.

Additionally, in the future we intend to combine the proposed method with others. We need to further examine speech enhancement and noise rejection and determine a way to remove overlapping noise. We also intend to study image recognition. Many researchers have added metadata to television broadcasting and combined two or more processes [2], [18]: for example, image enhancement combined with natural language processing. Previous studies have shown that image recognition used on characters' faces is quite effective. When television characters speak, their image is usually shown on the screen. Therefore, if characters in a scene can be identified by image recognition, it would make it easier to narrow down the target person in speaker identification.

References

[1] Photina Jaeyun Jang and A. G. Hauptmann, "Improving acoustic models with captioned multimedia speech", ICMCS-99 International Conference on Multimedia Computer Systems, pp. 767-771, 1999.

[2] H. Kuwano, Y. Matsuo, and K. Kawazoe, "Effects of Task-Cost Reduction on Metadata Generation Using Audio/Visual Indexing and Natural-Language Processing Techniques", The Institute of Image Information and Television Engineers, Vol. 61, pp. 842-852, 2007, (in Japanese).

[3] S. Meignier, D. Moraru, C. Fredouille, J. Bonastre, and L. Besacier, "Step-bystep and integrated approaches in broadcast news speaker diarization", Computer Speech and Language, Vol. 20, pp. 303-330, 2006.

[4] I. Yamada, M. Sano, H. Sumiyoshi, M. Shibata, and N. Yagi "Automatic generation of metadata for football games using announcer's commentary", IEIC Technical Report, VOL. 104, pp. 37-42, 2005, (in Japanese).

[5] A. Gordon, "Using annotated video as an information retrieval interface", International conference on intelligent user interface, New Orleans, LA, pp. 133-140, 2000.

[6] A. Messina, R. Borgotallo, G. Dimino, D. Airola Gnota, and L. Boch, "A Complete System for Automatic News Programme Annotation Based on Multimodal Analysis", Image Analysis for Multimedia Interactive Services, pp. 219-222, 2008.

[7] Chuck Wooters, "The ICSI RT07s Speaker Diarization System", Proceedings of the Second International Workshop on Classification of Events, Activities, and Relationships, Vol. 46, pp. 509-519, 2007.

[8] M. H. Kolekar, K. Palaniappan, and S. Sengupta, "A Novel Framework for Semantic Annotation of Soccer Sports Video Sequences", IET 5th European Conference on Visual Media Production, pp. 1-9, 2008.

[9] A. Baba, Y. Nishimoto, K. Ishikawa, H. Nakamura, T. Yoshimura, and T. Kurioka, "A Study on Metadata Technology for Broadcasting System based on Home Servers", IEIC Technical Report, VOL. 104, pp. 11-16, 2004, (in Japanese).

[10] "Results of broadcast captioning" Ministry of Internal Affairs and Communications, pp. 1-3, 2009, (in Japanese).

[11] R. Turetsky and N. Dimitrova, "Screenplay alignment for closed-system speaker identification and analysis of feature films", Proc. ICME, Vol. 3, pp. 1659-1662, 2004.

[12] T. Kawahara, A. Lee, K. Takeda, K. Itou, and K. Shikano, "Recent Progress of Open-Source LVCSR Engine Julius and Japanese Model Repository", Proc. ICSLP, Vol. 4, pp. 3069-3072, 2004.

[13] P. C. Woodland, J. J. Odell, V. Valtchev, and S. J. Young, "Large vocabulary continuous speech recognition using HTK", Proc. ICASSP, Vol. 2, pp. 125-128, 1994.

[14] M. Nishida and T. Kawahara, "Unsupervised Speaker Indexing Using Speaker Model Selection Based on Bayesian Information Criterion", IEIC Technical Report, vol. J87-D-II(2), pp. 504-512, 2004, (in Japanese).

[15] Z. Zhipeng and S. Furui, "Noisy Speech Recognition Based on Robust End-point Detection and Model Adaptation", Proc. ICASSP, Vol. 1, pp. 441-444, 2005.

[16] Shiro Ikeda, "Construction of Phone HMM Using Model Search Method", IEIC Technical Report, Vol. J78-D-2(1), pp. 10-18, 1995, (in Japanese).

[17] D. A. Reynolds and P. Torres-Carrasquillo, "Approaches and Applications of Audio Diarization", Proc. ICASSP, Vol. 5, pp. 18-23, 2005.

[18] K. Sekiguchi, M. Kosugi, and N. Mukai, "Automatic Indexing of the Baseball Game by Using Video and Audio Information", The Institute of Image Information and Television Engineers, Vol. 106, pp. 41-46, 2006, (in Japanese).

[19] T. Kosaka, T. Akatsu, and M. Katoh, "Speaker Vector-Based Speaker Identification with Phonetic Modeling", IEIC Technical Report, Vol. J90-D No. 12, pp. 3201-3209, 2007, (in Japanese).

 ISBN: 978-1-61208-157-1