SPEAKER INDEXING FOR NEWS ARTICLES, DEBATES AND DRAMA IN BROADCASTED TV PROGRAMS

M.Nishida and Y.Ariki

Department of Science and Technology, Ryukoku University
Oe-cho, Seta, Otsu-shi, Shiga, 520-2194, Japan
nishida@arikilab.elec.ryukoku.ac.jp, ariki@rins.ryukoku.ac.jp

Abstract

In this paper, we propose a method to extract and verify individual speaker utterance using a subspace method. This method can extract speech section of the same speaker by repeating speaker verification between the present speech section and the immediately previous speech section. The speaker models are automatically trained in the verification process without constructing speaker templates in advance. As a result, this speaker verification method is applied to speaker indexing. In this study, announcer utterances are automatically separated from news speech data which includes reporter or interviewer utterances using the speaker verification method. Also the utterances of each participant in debate program broadcasted on TV are automatically extracted. Furthermore, speech sections of actor or actress in TV drama are extracted.

1. Introduction

We are getting much information everyday from broadcasted TV programs such as news, debate, dramas, documents and so on. When we want to pick up some topics spoken by a certain person in debate or want to see some scenes where a certain actor plays, no VCR at present can search his speech and play back them.

In this paper, we propose a method to automatically divide the TV program speech into speakers and then index in real time who is speaking. Speaker models are not prepared in advance. They are constructed through indexing in self-organization mode. As a result, we can pick up the speech of the same person from the TV program.

In the speaker modeling, we employ a subspace method. Namely the speaker subspace of the first speaker is constructed using his input speech data. The speaker indexing is carried out based on speaker verification. Namely, for every spoken sentence, the input speech is verified whether it belongs to the same person just previously speaking.

If it belongs to the same person, then the speaker verification continues and his model is updated using the latest speech data. Otherwise the input speech is verified whether it belongs to one of the previous speakers. If so, the present speaker is regarded as the previous one. Otherwise a new speaker model is constructed using the following input speech data. This self-organized speaker indexing continues until the end of the TV program.

In the application to TV news program, it becomes possible to extract only the announcer speech, excluding the interviewer or reporter speech. Then the news can be reduced and summarized. In debate program, it becomes possible to construct a database about participant opinions by finding their utterances and dictating them automatically. Furthermore, in drama, actor or actress speech can be extracted and the corresponding scenes can be played back.

2. Related works

Several works have been reported about speaker indexing. They are mainly grouped into two classes; manual modeling and self-organized modeling. Almost all the works reported the manual modeling of speakers[1][2][3]. In these works, speech data of the individual speaker is collected in advance and is used to construct the speaker models. In speaker indexing, the input speech is compared with the speaker models and is accepted or rejected as the true speaker.

On the other hands, in self-organized modeling, the speaker models are constructed in on-line mode[4]. Namely speaker model construction and speaker indexing are performed simultaneously. Our proposing method is based on this self-organized modeling. The difference from the conventional method is that the proposing method can perform the speaker indexing in real time as well as in on-line mode. This means that speaker indexing and model construction can be performed simultaneously and sequentially without storing all the testing speech data in advance.

3. Speaker verification by subspace

3.1. Speaker verification

Speaker verification is a technique to judge if the input speech belongs to the specified person or not[5]. Fig.1 shows the speaker verification process. When the speaker ID of speaker A and his speech are fed to the verification system, the distance is computed between the model of the speaker A and the input speech. If the distance is smaller than some threshold, the input speaker is accepted as the true speaker A. Otherwise the input speaker is rejected. In our experiment, speaker subspace is constructed as the speaker model and the distance between the speaker subspace and the input speech is computed.

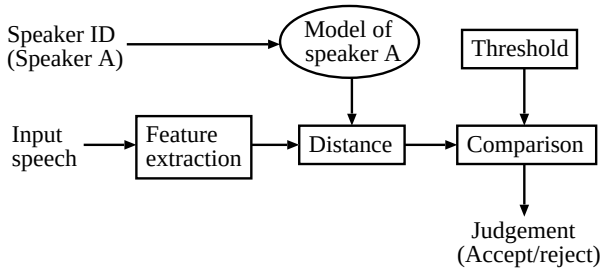

Figure 1: Speaker verification

3.2. Speaker subspace

As shown in Fig.2, we observe speech data $X^{(i)}$ of the speaker i and speech data $X^{(j)}$ of the speaker j in an observation space. The speech data are a sequence of spectral feature vectors $x_t^{(i)}$ and $x_t^{(j)}$ obtained at time t by short time spectral analysis. We denote the speech data $X^{(i)}$ as a matrix whose row is a spectral feature vector ${x_t^{(i)}}^T - {\mu^{(i)}}^T$ $(1 \leq t \leq M)$. Here $x_t^{(i)}$ denotes an observed feature vector and $\mu^{(i)}$ is their mean vector. The column of the matrix corresponds to frequency f $(1 \leq f \leq N)$.

By singular value decomposition, the speech data matrix $X^{(i)}$ is decomposed as

$$X^{(i)} = U^{(i)} \Sigma^{(i)} {V^{(i)}}^T \tag{1}$$

Here $U^{(i)}$ and $V^{(i)}$ are the matrices whose columns are eigenvectors of $X^{(i)}{X^{(i)}}^T$ and ${X^{(i)}}^T X^{(i)}$ respectively. $\Sigma^{(i)}$ is the singular value matrix of $X^{(i)}$.

The eigenvectors of the correlation matrix ${X^{(i)}}^T X^{(i)}$ are the orthonormal bases of the speech data $X^{(i)}$, computed based on a criterion that the total distance is minimized between feature vectors $x_t^{(i)} - \mu^{(i)}$ and the orthonormal bases[6][7]. Then $V^{(i)}$ is considered as orthonormal bases of the speaker space. This is completely same as the principal component analysis of the speech data $x_t^{(i)}$.

If the large singular values up to r numbers are selected from the matrix $\Sigma^{(i)}$, the matrix $V^{(i)}$ becomes N x r dimension and is considered as the speaker subspace[8].

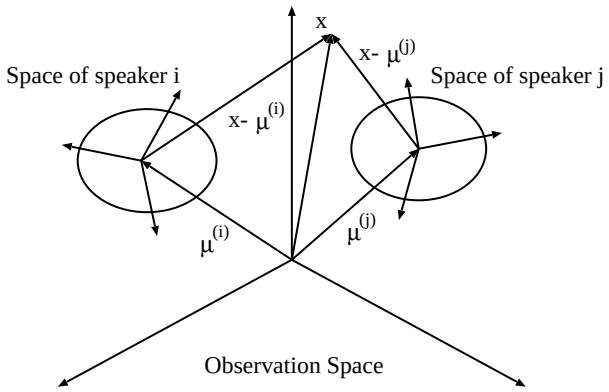

Figure 2: Speaker subspace

3.3. Verification by speaker subspace

The speaker subspace $V^{(i)}$ is composed of orthonormal bases $\{v_1^{(i)}, \cdots, v_r^{(i)}\}$ of the speech data $X^{(i)}$. Speaker verification can be carried out by computing a distance from an input speech vector x_t in the observation space to the speaker subspace $V^{(i)}$.

The distance is defined as follows using the orthonormal bases $\{v_1^{(i)}, \cdots, v_r^{(i)}\}$ from the observation space to the speaker subspace.

$$Dist(V^{(i)}, x_t) = \|x_t - \{\sum_j ((x_t - \mu^{(i)})^T v_j^{(i)}) v_j^{(i)} + \mu^{(i)}\}\|^2 \tag{2}$$

The distances computed by Eq.(2) between speech vectors x_t and the speaker subspace $V^{(i)}$ are averaged over time t. The speaker is identified as one with the minimum averaged distance between the speech vectors and the subspace.

4. Speaker indexing for TV news

4.1. Extraction of speaker section

Continuous news speech is divided into sections of respective speaker. The sections are called here " speaker sections". The continuous news speech is also divided into sections separated by silence. The sections are called " speech sections".

Fig.3 shows the concept of the speaker sections in TV news program. Our purpose is to extract the speech section of the same person from the total input speech. In the

figure, announcer speech is extracted and shown by the bold solid lines. The extraction process is shown in Fig.4 and summarized as follows;

(1) Averaged power is computed at every 1 second on the input speech. If it is lower than some threshold it is regarded as silence. The speech section between two silences is extracted.

(2) Using the firstly extracted speech section, a speaker subspace is constructed. This speaker subspace corresponds to the model of the speaker A shown in Fig.1. Here the threshold θ to accept or reject the speaker is determined as follows, using μ (mean) and σ (standard deviation) of the distance between speech data of the first speech section and the constructed speaker subspace.

$$\theta = \mu + \frac{\sigma}{3} \tag{3}$$

(3) On the successive speech section, the distance is computed between the input speech and the model. If the distance is lower than the threshold θ, it is judged that the speaker A is still speaking. In this case, the speaker subspace model is updated as well as the threshold θ using all the speech data verified as speaker A.

(4) Otherwise, it is regarded that speaker A has finished his speech and new speaker or previous speaker begins speaking. To judge it, the distance between the input speech section and the previously constructed speaker subspace models is computed. If some speakers have lower distance than threshold θ, then the input speaker is judged as the speaker with the lowest distance. Otherwise, the input speaker is regarded as a new speaker and step (2) begins starting.

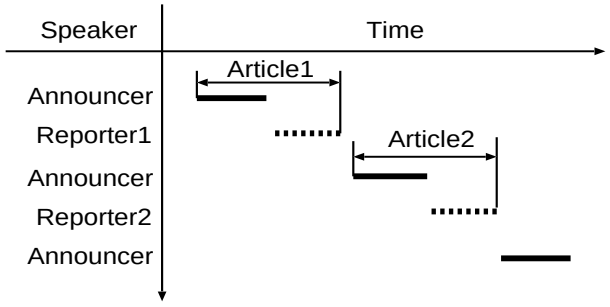

Figure 3: Extraction concept of speaker sections in TV news program

4.2. Experimental result

We selected 30 days 5 minutes NHK news articles which included reporter speech as well as announcer speech. The

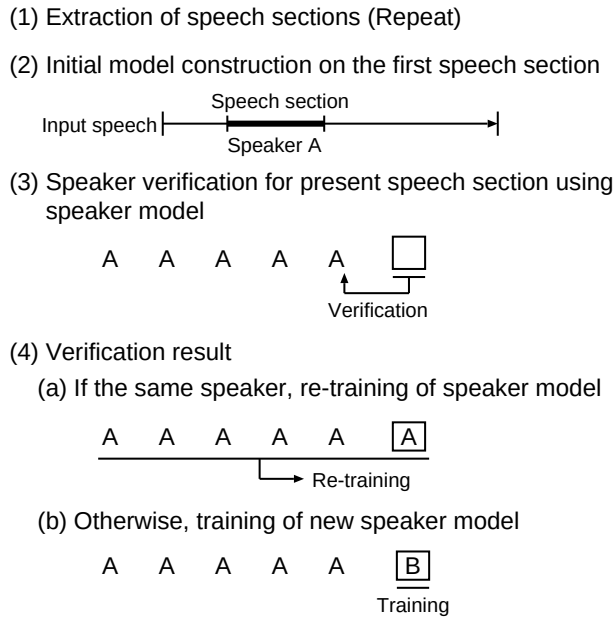

Figure 4: Extraction process of speaker sections in TV news program

duration time in total was about 150 minutes. For these 30 days news articles we carried out the experiment to extract the announcer sections. The dimension of speaker subspace was set to 7 after preliminary experiment. The experimental condition is shown in Table.1.

Table 1: Experimental condition

Speech data	30 days NHK news articles
Sampling frequency	12kHz
Frame length	20ms
Frame period	5ms
Window type	Hamming window
Features	LPC Cepstrum(16 orders)
Subspace dimension	7
Threshold θ	$\theta = \mu + \frac{\sigma}{3}$

The extraction of announcer sections was evaluated by the extraction rate and the precision rate defined as follows;

$$Extraction\ rate = \frac{\left\{\begin{array}{l}\text{Number of correctly verified speech}\\ \text{sections as announcer}\end{array}\right\}}{\left\{\begin{array}{l}\text{Number of total speech sections of}\\ \text{the announcer}\end{array}\right\}} \tag{4}$$

$$Precision\ rate = \frac{\left\{\begin{array}{l}\text{Number of correctly verified speech}\\ \text{sections as announcer}\end{array}\right\}}{\left\{\begin{array}{l}\text{Number of verified speech sections}\\ \text{as announcer}\end{array}\right\}} \tag{5}$$

Here announcer is judged as the speaker who speaks the longest time in 1 day 5 minutes NHK news.

The announcer extraction result is shown in Table2. The extraction rate was 93.4% and the precision rate was 98.7%. In a case where the first speech section was too short to construct the speaker subspace model, the threshold tended to be set lower than the optimal value so that the speech sections were sometimes rejected. In a case where the noise was superimposed on the speech, the speaker verification tended to fail.

Table 2: Speaker indexing result in news article(%)

Extraction rate	93.4
Precision rate	98.7

5. Speaker indexing for debate program

5.1. Extraction of speaker section

Continuous debating speech is divided into sections of respective speaker. The sections are called here " speaker sections". The continuous debating speech is also divided into sections separated by silence. The sections are called " speech sections".

The extraction process of speaker sections is similar to that in the news speech. Fig.5 shows the concept of speaker sections in the debate program. The same person, for example, speaker A is tracked and all his speech are extracted and shown by the solid lines. The difference of the extraction process from the news speech is summarized as follows;

1. Averaged power is computed at every 0.5 second in stead of 1 second on the input speech and silences are detected. This is because participants speak faster and silence is shorter between speakers in debate program compared with the news program.

2. If the first speech section is less than 1 second, it is neglected because it is too short for speaker model construction.

3. The threshold θ to accept or reject the speaker is set at two levels as follows;

   (a) In a case where the total speech data collected is less than 10 seconds, the threshold is set relatively higher as follows, because speech is changeable compared with normal speech in debate and the speaker model constructed by less than 10 seconds is unstable.

   $$\theta = \mu + 1.2\sigma \tag{6}$$

   (b) In a case where the total speech data collected is more than 10 seconds, the threshold is set relatively lower as follows, because the speaker model constructed by more than 10 seconds is stable.

   $$\theta = \mu + 0.65\sigma \tag{7}$$

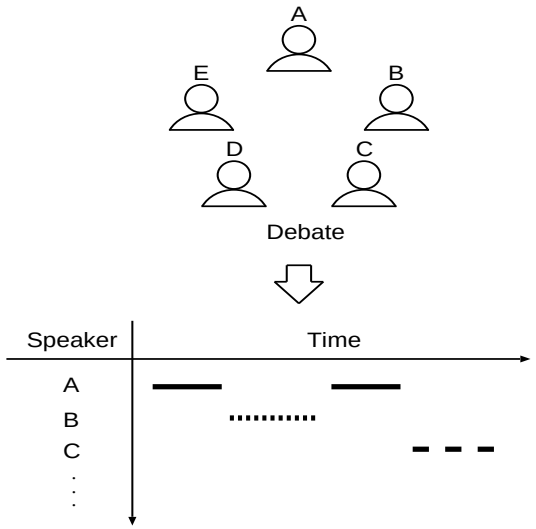

Figure 5: Extraction concept of speaker sections in debate program

5.2. Experimental result

TV video data in which five males were talking for 7 minutes was used for speaker indexing. The dimension of speaker subspace was set to 7 after preliminary experiment. The experimental condition is shown in Table.3.

Table 3: Experimental condition

Speech data	7 minutes debate program
Sampling frequency	12kHz
Frame length	20ms
Frame period	5ms
Window type	Hamming window
Features	LPC Cepstrum(16 orders)
Subspace dimension	7

The speaker indexing in debate was evaluated by the extraction rate and precision rate which are same as those in

the news speech. The speaker indexing result is shown in Table4. Here 1-θ indicates that the one threshold defined in the news article is used in this debate program. The 2-θ indicates the proposed two level thresholding.

In the 1-θ method, the extraction rate was 59.3% and the precision rate was 70.3% respectively. On the other hand, in the 2-θ method, the extraction rate was 96.7% and the precision rate was 100% respectively. This means that in a case where the training speech was too short to construct the speaker subspace model, the threshold tended to be set lower than the optimal value so that the speech was sometimes rejected. This problem was solved by utilizing the 2-θ method.

Table 4: Speaker indexing result in debate program(%)

	1-θ	2-θ
Extraction rate	59.3	96.7
Precision rate	70.3	100

6. Speaker indexing for drama

6.1. Extraction of speaker section

The extraction process for speaker sections in drama is almost same as that in the debate program except for the threshold. In drama, actors or actresses speak emotionally so that the distance deviation seems larger than that in the debate program. To deal with this problem, we employed 3-θ method by expanding 2-θ method used in the debate program.

The difference from the drama program is summarized as follows. The threshold θ to accept or reject the speaker is set at three levels as follows;

(a) In a case where the total speech data collected is less than 5 seconds, the threshold is set the highest as follows, partially because speech is changeable compared with normal speech and partially because the speaker model constructed by less than 5 seconds is unstable.

$$\theta = \mu + 1.6\sigma \tag{8}$$

(b) In a case where the total speech data collected is more than 5 seconds and less than 10 seconds, the threshold is set relatively higher as follows.

$$\theta = \mu + 1.4\sigma \tag{9}$$

(c) In a case where the total speech data collected is more than 10 seconds, the threshold is set relatively lower as follows because the speaker model becomes stable.

$$\theta = \mu + 0.7\sigma \tag{10}$$

6.2. Experimental result

TV drama in which three actors and two actresses were talking for 7 minutes was used for speaker indexing. At present, in this drama, the music nor strong noises are not included. The dimension of speaker subspace was set to 7 after preliminary experiment. The experimental condition is shown in Table.5.

Table 5: Experimental condition

Speech data	7 minutes drama including 5 persons
Sampling frequency	12kHz
Frame length	20ms
Frame period	5ms
Window type	Hamming window
Features	LPC Cepstrum(16 orders)
Subspace dimension	7

The speaker indexing in the drama was evaluated by the extraction rate and precision rate which are same as those in the debate program. The speaker indexing result is shown in Table6. Here 2-θ indicates that the two thresholds defined in the debate program are used in this drama. The 3-θ indicates the proposed three level thresholding.

Table 6: Speaker indexing result in drama(%)

	2-θ	3-θ
Extraction rate	58.7	61.9
Precision rate	68.3	69.8

In the 2-θ method, the extraction rate was 58.7% and the precision rate was 68.3% respectively. On the other hand, in the 3-θ method, the extraction rate was 61.9% and the precision rate was 69.8% respectively. This means that in a case where the training speech was too short to construct the speaker subspace model, the threshold tended to be set lower than the optimal value so that the speech was sometimes rejected. Longer the speech data are collected, the lower the threshold should be set. At present, we employed three level thresholding and obtained a little improvement. To solve this problem completely, we are planning a continuous thresholding method in which the threshold changes as a function of time.

7. Application of speaker indexing

7.1. Article extraction

We already applied this speaker indexing method to article extraction from TV news program. The TV news program is generally composed of announcer speech followed by reporter or interviewer speech and again announcer speech for the next news article. We applied the speaker indexing method and separated the announcer speech from the reporter or interviewer speech. In this application, a news article is extracted as a time section, starting from announcer speech followed by reporter or interviewer speech ending at announcer speech just before the next article. By indexing announcer speech, The extracted news articles are summarized and a news database can be constructed.

7.2. Article classification

In order to construct a news database with a function of video on demand (VOD), it is required to classify news articles into topics. We also constructed a system which can dictate news speech, extract keywords and classify news articles based on the extracted keywords. As an experiment, we compared the classification performance of news articles in two cases; dictating only the announcer utterances which are automatically extracted and dictating a whole speech which includes reporter or interviewer utterances. As a result, we found that it is sufficient to dictate only the announcer utterance in classifying the news articles and it contributes to reduce the processing time[9].

7.3. Speaker Retrieval

In debate program, the speaker retrieval is feasible. This contributes to play back and listen to the speech of the special spaker. If dictation is applied to the debate speech, the speaker retrieval with keyword locating function will be available. In drama, the speaker retrieval will be possible to pick up the speech of the special actor or actress whom a watcher is interested in among all the drama included in the database.

8. Conclusion

The method of real time speaker indexing has been proposed using subspace method.

It was applied to 30 days NHK news program in order to extract announcer speech. It was also applied to TV debate program in order to separate the speakers and retrieve them. Furthermore, speaker indexing was experimented for the TV drama program.

In the news program, the experiment showed 93.4% extraction rate and 98.7% precision rate. In the debate program, the experiment showed 96.7% extraction rate and 100% precision rate. However, in the drama program, the experiment showed relatively lower result, 61.9% extraction rate and 69.8% precision rate.

We are planning to improve the drama indexing and to make it robust under noisy and musical circumstances as well as speaker overlapping. We are also planning to expand this method to retrieve the speakers in debate and drama.

9. References

[1] L.Wilcox, F.Chen, D.Kimber and V.Balasubramanian, “SEGMENTATION OF SPEECH USING SPEAKER IDENTIFICATION”, ICASSP94, pp.161-164, 1994.

[2] A.E.Rosenberg, I.Magrin-Chagnolleau, et al, “SPEAKER DETECTION IN BROADCAST SPEECH DATABASES”, ICSLP98, pp.1339-1342, 1998.

[3] D. Roy and C. Malamud, “Speaker identification based test to audio alignment for an audio retrieval system”, ICASSP97, Munich, Vol. 2, pp. 1099-1103, 1997.

[4] D. Roy, “Speaker indexing using neural network clustering of vowel spectra”, International Journal of Speech Technology, Vol. 1, No. 2, pp.143-149, 1997.

[5] T.Matsui and S.Furui, “Comparison of text independent speaker recognition methods using VQ distortion and discrete/continuous HMMs”, Proc.ICASSP, Vol.II, pp157-160, 1992.

[6] Y.Ariki and K.Doi, “Speaker Recognition based on Subspace Method”, ICSLP94, pp.1859-1862, 1994.

[7] Y.Ariki, S.Tagashira and M.Nishijima, “Speaker Recognition and Speaker Normalization by Projection to Speaker Subspace”, ICASSP96, sp9.1, pp.319-322, 1996.

[8] E.Oja, “Subspace Methods of Pattern Recognition”, Research Studies Press, England, 1983.

[9] Y.Ariki, J.Ogata and M.Nishida, “News Dictation and Article Classification Using Automatically Extracted Announcer Utterance”, AMCP98, pp.78-89, 1998.