# REAL TIME SPEAKER INDEXING BASED ON SUBSPACE METHOD

## - APPLICATION TO TV NEWS ARTICLES AND DEBATE -

*M.Nishida and Y.Ariki*

Department of Electronics and Informatics
Ryukoku University, Seta, Otsu-shi, Shiga, 520-2194, Japan
nishida@arikilab.elec.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 extracted from news speech data which includes reporter or interviewer utterances. Also extracted automatically are the utterances of each participator in debate program broadcasted on TV.

## 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 as in debate, 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 sentences, 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 are reduced and summarized. In debate program, it becomes possible to construct a database about participators opinions by finding their utterances and dictating them automatically.

## 2. SPEAKER VERIFICATION BY SUBSPACE

### 2.1. Speaker Verification

Speaker verification is a technique to judge if the input speech belongs to the specified person or not[1]. 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 input speech and the speaker subspace is computed.

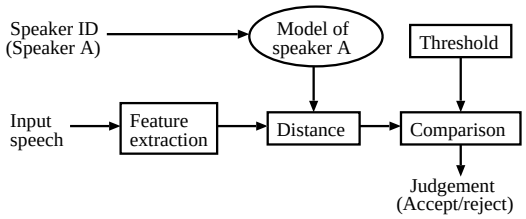


**Figure 1**: Speaker verification

### 2.2. 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 ob-

servation 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 $x_t^{(i)} - \mu^{(i)}$ $(1 \leq t \leq M)$. 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[2][3]. 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[4].

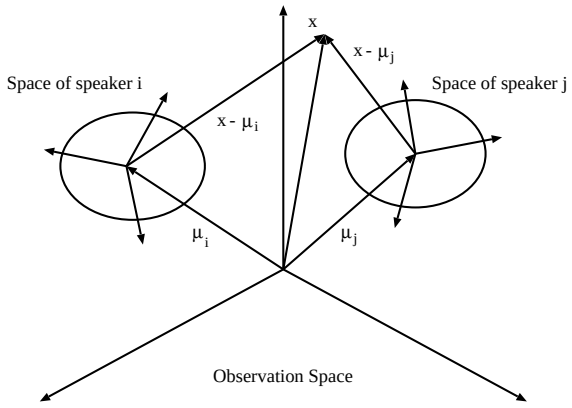


**Figure 2**: Speaker subspace

## 2.3. 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 presented as follows using the orthonormal bases $\{v_1^{(i)}, \cdots, v_r^{(i)}\}$ from the observation space to the speaker subspace.

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

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.

# 3. SPEAKER INDEXING FOR NEWS ARTICLES

## 3.1. 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". The extraction process for speaker sections is 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 $\theta$ to accept or reject the speaker is determined as follows, using $\mu$ (mean) and $\sigma$ (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 $\theta$, it is judged that the speaker $A$ is still speaking. In this case, the speaker subspace model is updated as well as the threshold $\theta$ 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 $\theta$, 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.

### 3.2. Experimental Result

We selected 30 days 5 minutes NHK news articles which included reporter speech as well as announcer speech. The 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 $\theta$ | $\theta = \mu + \frac{\sigma}{3}$ |

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

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

$$\text{Precision rate} = \frac{\left\{\begin{array}{l}\text{Number of correctly verified}\\ \text{speech sections as announcer}\end{array}\right\}}{\left\{\begin{array}{l}\text{Number of verified speech sec-}\\ \text{tions as announcer}\end{array}\right\}} \quad (5)$$

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**: Experimental result(%)

| | |
|---|---|
| Extraction rate | 93.4 |
| Precision rate | 98.7 |

## 4. SPEAKER INDEXING FOR DEBATE

### 4.1. 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"[5].

The extraction process of speaker sections is similar to that for the news speech. However, in debate, sometimes speakers spoke overlappingly so that speaker verification is carried out at every 0.5 second in stead of every speech section. Further more, in debate, speech is changeable compared with normal speech so that the threshold to accept or reject is set higher than that for news speech. In order to detect the speaker change safely, successive 3 times failures of the speaker verification causes the speaker change. The extraction process is as follows;

(1) Averaged power is computed at every 0.5 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 extracted speech section with more than 5 seconds, a speaker subspace is constructed. This speaker subspace corresponds to the model of the speaker $A$ shown in Fig.1. Here the threshold $\theta$ to accept or reject the speaker is determined as follows, using $\mu$ (mean) and $\sigma$ (standard deviation) of the distance between speech data of the first 5 seconds speech section and the constructed speaker subspace.

$$\theta = \mu + \frac{\sigma}{2} \quad (6)$$

(3) On the successive speech, at every 0.5 second, the distance is computed and averaged between the input speech and the model. If the distance is lower than the threshold $\theta$, it is judged that the speaker $A$ is still speaking. In this case, the speaker subspace model is updated as well as the threshold $\theta$ using all the speech data verified as speaker $A$.

(4) Otherwise, if the distance is higher than the threshold $\theta$ for successive 3 times, 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 and the previously constructed speaker subspace models are computed. If some speakers have lower distance than threshold $\theta$, 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.

## 4.2. Experimental Result

TV video data in which five persons were talking for 7 minutes was used for speaker indexing. The dimension of speaker subspace was set to 5 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 | 5 |
| Threshold $\theta$ | $\theta = \mu + \frac{\sigma}{2}$ |

The evaluation of the speaker indexing was carried out by verification rate, recall rate and precision rate which are defined as follows;

$$\text{Verification rate} = \frac{\text{Number of the correct verification}}{\text{Number of verification at every 0.5 second}} \quad (7)$$

$$\text{Recall rate} = \frac{\text{Number of correctly verified boundaries}}{\text{Number of speaker boundaried}} \quad (8)$$

$$\text{Precision rate} = \frac{\text{Number of correctly verified boundaries}}{\text{Number of extracted speaker boundaries}} \quad (9)$$

The speaker indexing result is shown in Table4. The verification rate was 95.5% high. However, the recall rate and precision rate were 75.0% and 60.0% respectively. 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. Since in this experiment, we used 5 seconds speech for training and obtained 95.5% high speaker verification rate, it can be said that 5 seconds training speech was enough.

In a case where the true speaker is verified as the different person in the course of the verification, the speech is segmented and the false speaker section occurs. On the other hand, in a case where the different speaker is verified as the same speaker, the re-training of the speaker model produces the wrong speaker model so that the speaker boundaries tend to be wrong hereafter. In this way, the recall and precision rate became low in spite of the high verification rate.

**Table 4**: Experimental result

| | Number of verification | % |
|---|---|---|
| Verification rate | 640 / 670 | 95.5 |
| | Number of boundaries | % |
| Recall rate | 6 / 8 | 75.0 |
| Precision rate | 6 / 10 | 60.0 |

## 5. 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.

In the debate program, the experiment showed 75.0% recall rate and 60.0% precision rate. In the news program, the experiment showed 93.4% recall rate and 98.7% precision rate. The reason why the recall and precision rate are high in the news program compared to the debate program is that the speaker verification was carried out for each sentence extracted using the power level instead of verification at every 0.5 second, according to the fact there are few speech overlaps between the different speakers in the news program.

We are planning to apply this method to news summary including dictation as well as to drama or sports news.

## 6. REFERENCES


1. 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.

2. Y.Ariki and K.Doi, " Speaker Recognition based on Subspace Method", ICSLP'94, pp.1859-1862, 1994.

3. Y.Ariki, S.Tagashira and M.Nishijima, " Speaker Recognition and Speaker Normalization by Projection to Speaker Subspace", ICASSP'96, sp9.1, pp.319-322, 1996.

4. E.Oja," Subspace Methods of Pattern Recognition", Research Studies Press, England, 1983.

5. Y.Sugiyama, N.Ishikawa, M.Nishida and Y.Ariki, " Indexing and Retrieval of Human Individuals on Video Data Using Face and Speaker Recognition", IWAIT'98, pp.122-127, 1998.