# Unsupervised Speaker Indexing using Anchor Models and Automatic Transcription of Discussions

*Yuya Akita*[†‡] *Tatsuya Kawahara*[†‡]

[†] School of Informatics, Kyoto University
Kyoto 606-8501, Japan
[‡] Japan Science and Technology Corporation, PRESTO

## Abstract

We present unsupervised speaker indexing combined with automatic speech recognition (ASR) for speech archives such as discussions. Our proposed indexing method is based on anchor models, by which we define a feature vector based on the similarity with speakers of a large scale speech database. Several techniques are introduced to improve discriminant ability. ASR is performed using the results of this indexing. No discussion corpus is available to train acoustic and language models. So we applied the speaker adaptation technique to the baseline acoustic model based on the indexing. We also constructed a language model by merging two models that cover different linguistic features. We achieved the speaker indexing accuracy of 93% and the significant improvement of ASR for real discussion data.

## 1. Introduction

In recent years, digital archives of speech materials have come to be available. For quick browsing of such archives, indices are quite useful and therefore they are an essential part of archives. In this paper, we present a method of speaker indexing of discussions, in which several speakers make utterances and speaker labels are important indices. We also address automatic transcription, which leads to topic indices.

Speaker indexing should be performed in an unsupervised manner. Supervised training of speaker models, which is commonly used for speaker identification, is not practical because participants often change in discussions. So, we propose a method of unsupervised indexing that uses only the discussions to be indexed.

For accurate transcription, ASR system needs dedicated acoustic and language models to the task. We perform unsupervised adaptation of acoustic model using the speaker indexing result. We also investigate possible solutions for adequate language model.

## 2. Speaker indexing based on anchor models

### 2.1. Feature projection using anchor models

With the conventional unsupervised method (such as [1]), which incrementally cluster the utterances, the number of speaker clusters increases with time, thus a huge number of clusters are generated for long speech like those in discussions. It is not easy to determine whether a new utterance is made by a new speaker or by someone who has already spoken. Thus, we introduce off-line indexing, in which whole segments of speech can be used for globally optimal speaker clustering.

There are some studies on off-line indexing using ergodic HMM[2, 3], which directly deals with acoustic features. Clustering results, however, are sensitive to variations in acoustic features. Actually, the approach realized limited performance. Another study uses cluster analysis for off-line indexing[4]. Every utterance is modeled by Gaussian Mixture Model (GMM), and utterances are aggregated based on a distance measure defined by the GMM. Duration of the test-set utterances is one minute in [4], while there are much shorter utterances such as those of a few seconds in real discussions. Therefore, reliable training of GMM is difficult for discussion materials.

In this paper, we introduce speaker features using anchor models[5]. Anchor models are GMMs trained for a pool of speakers, and a speaker characterization vector (SCV) is composed of a set of speaker recognition likelihood. The projection of acoustic features is based on matching with the statistical model, and expected to suppress variations in acoustic features, especially in spontaneous speech, while preserving differences of speaker characteristics.

Previous studies[5, 6] use SCV as a kind of speaker models in speaker detection and identification tasks. In both cases, absolute performance of SCV is not sufficient for the tasks. Actually in [5], SCV is used only for candidate selection of speaker identification, in which speaker models are trained with supervised training. In this study, SCV is used for clustering speakers, and we construct speaker models for the clusters, which are then used for indexing. Compared with previous works, no supervised training is involved. However, we found the simple application of anchor models led to only poor performance in achieving separability in clustering. Therefore, we incorporate several methods to extract discriminant features for speaker clustering. Thus, our proposed method consists of three steps; (1) enhancement of speaker separability of SCV, (2) unsupervised SCV clustering and (3) autonomous construction of speaker models and reidentification.

Table 1: Number of speakers and utterances in test-set discussions

| ID | 0624 | 0805 | 0819 | 0902 | 0916 | 1118 | 1125 | 1209 | 1216 | 0113 |
|---|---|---|---|---|---|---|---|---|---|---|
| #Speakers | 5 | 5 | 5 | 8 | 6 | 8 | 5 | 5 | 5 | 5 |
| #Utterances | 534 | 665 | 609 | 541 | 612 | 474 | 371 | 613 | 559 | 524 |

### 2.2. Discussion corpus

We compiled a discussion corpus using a TV program "Sunday Discussion" broadcasted by NHK (Japan Broadcasting Corporation). This program shows discussions on current topics in politics and economy by politicians, economists and experts in the fields. A chairperson also takes part and prompts the speakers. Utterances generally do not overlap, but there are short responses such as "yes" and "I see" as well as coughing and laughing. We did not remove such overlapping utterances. The speech was segmented into utterances based on detection of short pauses longer than 400 milliseconds. Duration of one discussion is one hour. Ten discussions are used for the experiments. The number of speakers and utterances in each discussion are shown in Table 1. The average number of utterances is 550.

## 3. Speaker indexing method

In this section, the proposed speaker indexing method is described. First, a set of anchor models are trained using a speech database. Secondly, anchor model selection, SCV normalization and dimensional reduction are performed to enhance SCV performance. Then the reduced SCVs are clustered using the LBG algorithm. In this study, we assume that the number of clusters (i.e., speakers) is given beforehand. Finally, speaker models (GMM) are constructed for every cluster, and speaker identification is performed using these GMMs. These steps are explained in the followings.

### 3.1. Training and selection of anchor models

Anchor models are a set of GMMs trained by using a large speech database. We used the ASJ-JNAS speech database, which is widely used to construct speaker-independent Japanese phone models for ASR, and is considered to be sufficient for constructing the SCV. To suppress the linguistic bias, only phoneme-balanced sentences are used. MFCC, $\Delta$MFCC and $\Delta$energy are used as acoustic features. The number of Gaussians in each GMM is 32. In total, 304 (153 male and 151 female) models are trained.

For accurate clustering, a number of anchor models that hardly match actual input speech should be eliminated. In [6], speaker models are clustered and merged for better representation of the speaker space, but it does not consider the input speech.

We reduce the anchor models depending on the input speech to be indexed. As a measure of reduction, the average normalized frame-wise score (computed in a manner described below for frames instead of utterances) is calculated for each anchor model using whole speech data. The models that give better scores as a whole are regarded as useful and selected.

We made preliminary experiments by changing the number $n$ of selected anchor models ($n = 50, 100, 150, 200, 250$ and original 304). The case of $n = 100$ showed the best performance.

### 3.2. Normalization of SCV

To compensate differences between the recording environments of the training and input speech, cepstral mean normalization is applied. The cepstral mean is calculated not by a single utterance but by the whole input to avoid losing speaker characteristics.

SCVs are calculated by speaker identification using anchor models. However, the absolute value of each component in an SCV varies because of factors other than speaker features. Since the proportion of SCV components is more important rather than their absolute values, we normalize every component $p_i$ of SCV $(p_1, p_2, \ldots, p_N)$ so that the distribution of these components has a mean of 0 and a variance of 1. The normalized SCV is characterized by the following:

$$p'_i = \frac{p_i - \bar{p}}{\sqrt{\frac{\sum_{i=1}^{N}(p_i - \bar{p})^2}{N}}} \qquad (1)$$

where $\bar{p}$ is the mean of $p_i$, and $N$ is the number of anchor models (304).

### 3.3. Feature extraction with KL transformation

Even after anchor model selection is performed, selected models (i.e. components of the SCV) are not necessarily useful. Several components are relatively constant and such components do not contribute to discrimination. To extract only discriminant features and remove these useless components, we perform Karhunen-Loève transformation (KLT) on the SCV.

As a threshold for the transformation, we adopt cumulative contribution (CC). Determining an optimal threshold value for KLT is difficult because each discussion has its own CC curve and the operating points are slightly different. We found that the best number of discriminant dimensions after KLT is small, and almost same as that of speakers. In the experiment, therefore the threshold on the CC is set to 0.7, which usually generate that range of dimensions.

### 3.4. Clustering and re-identification

The reduced SCV is clustered up to the number of speakers using LBG algorithm. Then, a speaker model of GMM is trained for every cluster using utterances in the speaker cluster. Acoustic features of the GMM are same as anchor models. Using these GMMs, input utterances are identified and re-labeled.

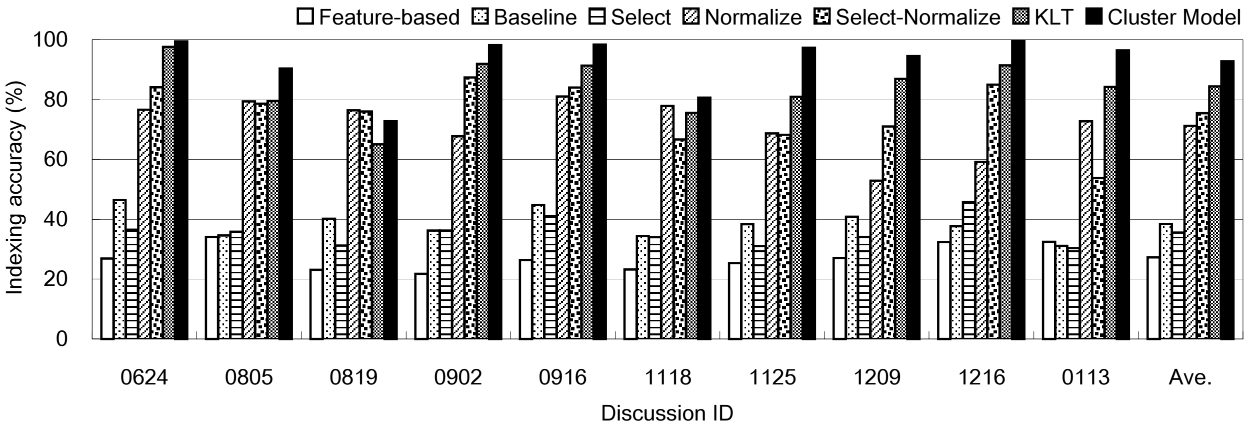

Figure 1: Speaker indexing result

### 3.5. Speaker indexing results

We made experimental evaluation on the proposed method using ten discussions described in Section 2.2.

As a measure of evaluation, we define speaker indexing accuracy for discussions. For every correspondence between the clusters $\{C_i\}$ and the speakers $\{S_j\}$, the number $(n_{ij})$ of $S_j$'s utterances classified into $C_i$ is calculated. Let $U$ is the total number of utterances, $L$ is the number of speakers (i.e., clusters) and $A(a_1, a_2, \ldots, a_L)$ is a set of assignments between cluster $C_i$ and speaker $a_i$. Then, accuracy of an assignment $s(A)$ is defined as $s(A) = \frac{1}{U}\sum_i^L n_{ia_i}$. Choosing the best assignment $A_{max}(= \arg\max_A s(A))$, the indexing accuracy is defined as $s(A_{max})$, which ranges from 0 at the worst to 1 at the best.

Figure 1 shows indexing results for each discussion. "Feature-based" is the accuracy of clustering with the acoustic feature vectors that consist of MFCC, $\Delta$MFCC and $\Delta$energy. It was only 27.3% because of the huge variation of these features. "Baseline" shows the accuracy of clustering with the original SCV. Although they showed better performance than the acoustic feature-based case, the accuracy was still very low.

"Select", "Normalize", "Select-Normalize" and "KLT" shows the accuracies obtained after anchor model selection, normalization of the original SCVs, normalization of the reduced SCVs, application of KL transformation to the reduced and normalized SCVs, respectively. Incorporation of three techniques drastically improved the accuracy. It suggests that appropriate handling of the SCV is vital in improving speaker separability. "Cluster Model" shows the final accuracy of indexing after identification with the models derived from clustering. The step further improves the accuracy to 92.7% finally.

## 4. Speech recognition of discussions

Next, we address automatic transcription of the discussions by making use of the speaker indexing result.

### 4.1. Language model

In "Sunday Discussion", we observe two kinds of linguistic features: (1) words and phrases on politics, economy and current topics, and (2) fillers and expressions peculiar to spontaneous speech. There is no text corpus containing plenty of these linguistic features to train a matched model for discussions.

Therefore, we construct a language model by merging two models representing above (1) and (2), respectively. As for (1), we train a newspaper model which contains political and economic topics. As for (2), we train a lecture model with "the Corpus of Spontaneous Japanese" (CSJ)[7], which consists of many lectures. We construct another model from the minutes of the National Diet of Japan. Table 2 shows details of these models. Test-set perplexity (PP) and out-of-vocabulary (OOV) rate (in Table 2) are calculated with the test-set discussions.

We made comparison experiments on merging these three models. Models were constructed using all possible combinations of the two or three of them, and we evaluated them with PP and OOV rate. Table 3 shows the result. The N+L+M model achieves the minimum PP and OOV rate among these models, and the L+M model gets comparable performance, since the minutes model covers topic words as well as the newspaper model, and the newspaper model does not contain spoken expressions. We could reduce both PP and OOV rate remarkably from either of the three models.

### 4.2. Speaker adaptation of acoustic model

Since there is no large speech database of discussions, a task-dependent acoustic model cannot be trained, either. In discussions, particular phenomena in spontaneous speech such as fast speaking and pronunciation variation occur. They are often observed in lectures similarly. Therefore, we adopt the acoustic model trained with lecture speech (60 hours) of the CSJ[8] as a baseline. It is a phonetic tied-mixture (PTM) triphone HMM with 16K Gaussians in total.

For this baseline model, unsupervised MLLR speaker adaptation is performed using the result of speaker index-

Table 2: List of language model

| | Newspaper (N) | Lecture (L) | Minutes (M) |
|---|---|---|---|
| Corpus | The Mainichi Newspaper (2001 version) | Corpus of Spontaneous Japanese | Minutes of the Japanese Diet |
| #Words | 21.7M | 2.7M | 64.1M |
| Vocab. size | 30K | 20K | 30K |
| Ave. PP | 347.42 | 223.89 | 207.54 |
| Ave. OOV | 5.36% | 5.15% | 5.51% |

Table 3: Perplexity (PP) and Out-Of-Vocabulary (OOV) rate by combined models

| | N+L | N+M | L+M | N+L+M |
|---|---|---|---|---|
| Vocab. size | 35K | 39K | 36K | 43K |
| Ave. PP | 195.94 | 218.18 | 152.13 | 149.34 |
| Ave. OOV | 2.52% | 4.44% | 2.30% | 2.11% |

Refer N, L and M to Table 2.

ing. For each participant, utterances that are labeled as the speaker are used for adaptation. As for phone transcriptions of utterances, the initial ASR result with the baseline acoustic model is used. The number of clusters in MLLR adaptation is 32.

For reference, we performed semi-supervised adaptation of the baseline model using correct speaker labels. (Transcriptions are given by the initial ASR.) We also tested the case of supervised adaptation using the correct speaker labels and manually transcribed text.

### 4.3. Speech recognition results

The ASR experiments are done using our decoder Julius 3.3[9]. Sequential decoding without prior segmentation is applied to deal with long (more than one minute) utterances. Figure 2 shows the word accuracy.

With the baseline lecture model, the accuracy was 51.0% on the average. The unsupervised speaker adaptation improved it to 56.9%. The figure is comparable to those of semi-supervised adaptation (57.3%) and supervised adaptation (58.9%). The result demonstrates that speaker adaptation based on the unsupervised speaker indexing is effective.

The recognition performance for discussions is lower than that for lectures[8], since acoustic and language models are not completely matched to the discussions, while models for lectures are trained with the lecture corpus (CSJ).

## 5. Conclusion

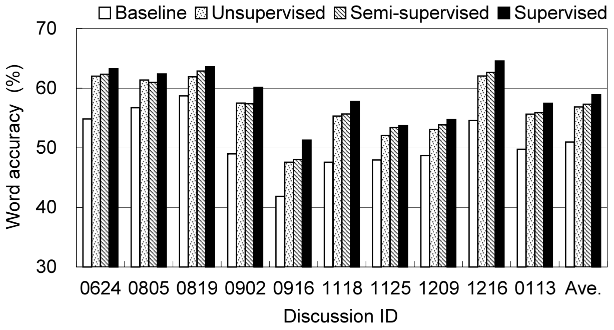

Figure 2: Speech recognition result (word accuracy)

We have proposed a method of unsupervised speaker indexing based on anchor models for long speech archives such as discussions. Speaker features are represented based on similarities between the input speech and those of many speakers using anchor models. The vector is normalized and reduced to suppress acoustic variations and extract discriminant features adaptively to the given input speech. These vectors are clustered and speaker models are autonomously trained for final indexing.

It is demonstrated that the vector normalization and reduction are effective in clustering, and that the completely unsupervised indexing method achieves the accuracy of 93% for real discussions.

We have also addressed automatic transcription of discussions using acoustic and language models trained without a matched corpus. Unsupervised adaptation of the baseline acoustic model is made possible by the speaker indexing, and it is shown to be effective. A language model is constructed by merging models representing different linguistic features. The overall framework effectively combines speaker indexing and speech recognition, and is realized in an unsupervised manner.

**Acknowledgment:** The authors are grateful to Prof. H. G. Okuno for his fruitful comments.

## 6. References

[1] M. Nishida and Y. Ariki. Real Time Speaker Indexing Based on Subspace Method - Application to TV News Articles and Debate. In *Proc. ICSLP*, 1998.

[2] J. Murakami, M. Sugiyama, and H. Watanabe. Unknown-Multiple Signal Source Clustering Problem Using Ergodic HMM and Applied to Speaker Classification. In *Proc. ICSLP*, 1996.

[3] J. Ajmera, H. Bourlard, I. Lapidot, and I. A. McCowan. Unknown-Multiple Speaker Clustering Using HMM. In *Proc. ICSLP*, 2002.

[4] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish. Clustering Speakers by their Voices. In *Proc. ICASSP*, 1998.

[5] D. Sturim, D. Reynolds, E. Singer, and J. Campbell. Speaker Indexing in Large Audio Databases Using Anchor Models. In *Proc. ICASSP*, 2001.

[6] Y. Mami and D. Charlet. Speaker Identification by Location in an Optimal Space of Anchor Models. In *Proc. ICSLP*, 2002.

[7] S. Furui, K. Maekawa, and H. Isahara. Toward the Realization of Spontaneous Speech Recognition – Introduction of a Japanese Priority Program and Preliminary Results –. In *Proc. ICSLP*, 2000.

[8] H. Nanjo and T. Kawahara. Speaking-rate Dependent Decoding and Adaptation for Spontaneous Lecture Speech Recognition. In *Proc. ICASSP*, 2002.

[9] A. Lee, T. Kawahara, and K. Shikano. Julius — an Open Source Real-Time Large Vocabulary Recognition Engine. In *Proc. EUROSPEECH*, 2001.