# AUTOMATIC TRANSCRIPTION OF DISCUSSIONS USING UNSUPERVISED SPEAKER INDEXING

*Yuya Akita*[†‡] *Masafumi Nishida*[‡] *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, and we incorporate several techniques 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 word recognition accuracy of 57% 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 in 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, 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[1, 2], which directly deals with acoustic features. Clustering results, however, is sensitive to variations in acoustic features. Actually, the approach realized limited performance. In this paper, we introduce speaker features using anchor models[3]. An anchor model is a Gaussian Mixture Model (GMM), and a speaker characterization vector (SCV) is composed of a set of likelihood. The projection of acoustic features is based on matching with the statistical models, and expected to suppress variations in acoustic features, especially in spontaneous speech, while preserving differences of speaker characteristics.

We found the simple application leads to only poor performance in this task. Therefore, we incorporate several methods to extract discriminant features for speaker clustering.

### 2.2. Indexing Processes

Flow of the proposed indexing method is shown in Figure 1. Anchor models are preliminarily trained by using a large speech database. We adopt 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. The number of anchor models (i.e. speakers in the database) is 304. To suppress the linguistic bias, only phoneme-balanced sentences are used.

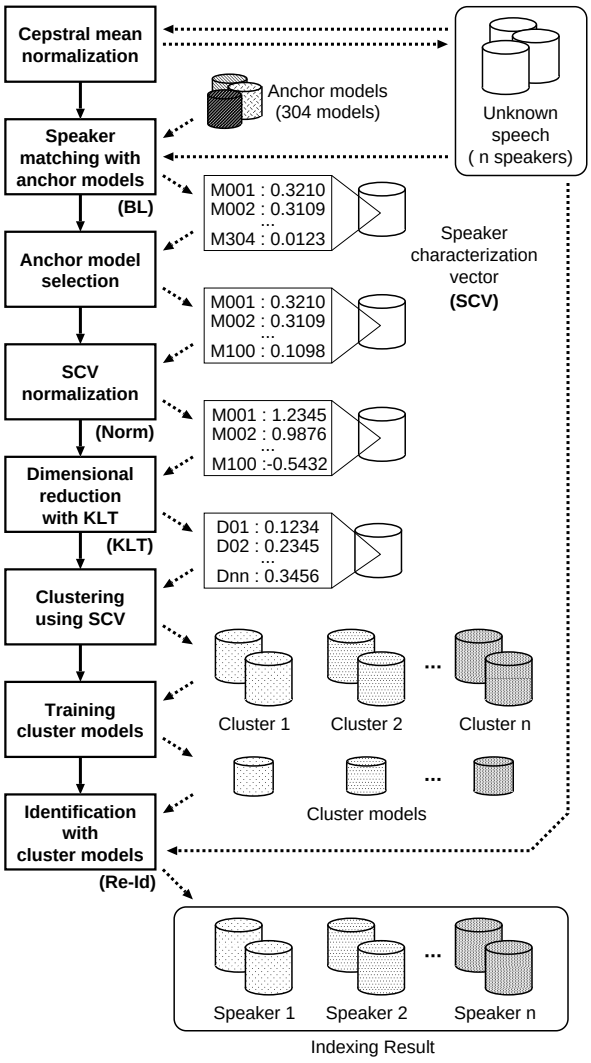


**Fig. 1**. Process flow of proposed speaker indexing method

First, we perform recording channel compensation by cepstral mean normalization (CMN). Secondly, speaker matching with anchor models is performed to generate SCVs. Here, to enhance the speaker separability, we introduce model selection that eliminates irrelevant anchor models depending on the input speech to be indexed. In [4], speaker models are clustered and merged for better representation of the speaker space, but it does not consider the input speech. As a measure of reduction, the average normalized frame-wise score (computed for frames instead of utterances) is calculated for each anchor model using whole speech frames, and 100 models are selected.

Then, normalization of the SCV is performed. The magnitude of components in the SCV varies because of factors other than speaker characteristics. Since the proportion of SCV components is more important rather than their magnitude, we normalize every component $p_i$ of SCV $V = (p_1, p_2, \ldots, p_N)$ so that the distribution of these components has a mean of 0 and a variance of 1.

**Table 1**. Speaker indexing result

| ID | #Spkr | #Uttr | BL | Norm | KLT | Re-Id |
|---|---|---|---|---|---|---|
| 0624 | 5 | 534 | 0.464 | 0.841 | 0.976 | 0.994 |
| 0805 | 5 | 665 | 0.346 | 0.786 | 0.795 | 0.904 |
| 0819 | 5 | 609 | 0.402 | 0.760 | 0.650 | 0.727 |
| 0902 | 8 | 541 | 0.362 | 0.874 | 0.919 | 0.982 |
| 0916 | 6 | 612 | 0.448 | 0.840 | 0.913 | 0.984 |
| 1118 | 8 | 474 | 0.344 | 0.667 | 0.755 | 0.806 |
| 1125 | 5 | 371 | 0.383 | 0.682 | 0.809 | 0.973 |
| 1209 | 5 | 613 | 0.408 | 0.710 | 0.869 | 0.945 |
| 1216 | 5 | 559 | 0.377 | 0.850 | 0.914 | 0.996 |
| 0113 | 5 | 524 | 0.311 | 0.538 | 0.842 | 0.964 |
| Average | 5.7 | 550 | 0.385 | 0.755 | 0.844 | 0.927 |

Refer BL, Norm, KLT and Re-Id to Fig. 1.

Most components of the SCV are nearly 0, and such components do not contribute to discrimination. To extract only discriminant features and remove the useless components, we perform Karhunen-Loève transformation (KLT) on the SCV.

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.

### 2.3. Discussion Corpus

We constructed 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 statesmen, economists and experts in the fields. A chairperson also takes part in the discussion 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. The total length of speech in one discussion is one hour. The average number of utterances is about 550. Ten discussions are used for the experiments.

### 2.4. Speaker Indexing Results

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 of $S_j$'s utterances classified into $\{C_i\}$ $(n_{ij})$ is calcu-

lated. Let $U$ is the total number of utterances, $L$ is the number of speakers (i.e., clusters) and $A(a_1, a_2, \ldots, a_L)$, $\{a_i\} = \{1, 2, \ldots, L\}$ 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.

Table 1 shows indexing results for each discussion. "BL" column shows the accuracy of clustering with the original SCV. And it was only 38.5%.

"Norm" shows the accuracy obtained with the normalized SCV after model selection. Incorporation of this technique drastically improves the accuracy. It suggests that appropriate handling of the SCV is vital to improve speaker separability. "KLT" shows the accuracy obtained after application of KL transformation. The feature extraction also has some effect. "Re-Id" shows the final accuracy of indexing after identification with the models derived from clustering. The step further improves the accuracy to 92.7% finally.

## 3. AUTOMATIC SPEECH RECOGNITION OF DISCUSSIONS

### 3.1. 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 for a matched model to these 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 "Corpus of Spontaneous Japanese" (CSJ)[5], 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 transcriptions of ten "Sunday Discussion" programs described in Section 2.3. The cut-off parameter of n-gram entries is set to 1 in all models.

We made preliminary 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 minimum PP and OOV rate among these models, and the L+M model showed comparable performance, since the minutes model covers topic words as well as the newspaper model, and the newspaper model does not contain spoken expressions.

**Table 2**. 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 for 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.

Thus, weighted merging of lecture-based and minutes-based models are done to set up the language model for discussions. The vocabulary size is 36,053. The weight ratio is determined as 0.5:0.5. Table 4 shows PP and OOV rate for each discussion. Average PP and OOV rate are 152.13 and 2.30%, respectively. We could reduce both PP and OOV rate remarkably from any of the three models.

### 3.2. 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 hesitations and pronunciation variations occur. They are often observed in lecture speech similarly. Therefore, we adopt the acoustic model trained with lecture speech in CSJ[6] as a baseline. In fact, the lecture model achieved better performance than a read-speech model. It is a PTM triphone HMM and its specifications are shown in Table 5.

For this baseline model, unsupervised MLLR speaker adaptation is performed using the result of speaker indexing . For each participant, utterances that are labelled 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 also perform supervised adaptation of the baseline model using correct speaker labels and manually transcribed text.

**Table 4**. Perplexity (PP) and Out-Of-Vocabulary (OOV) rate for test-set discussions

| ID | 0624 | 0805 | 0819 | 0902 | 0916 |
|---|---|---|---|---|---|
| PP | 127.94 | 142.97 | 143.09 | 161.58 | 225.26 |
| OOV | 1.70% | 1.97% | 2.00% | 2.20% | 2.91% |
| ID | 1118 | 1125 | 1209 | 1216 | 0113 |
| PP | 144.49 | 112.87 | 176.53 | 126.37 | 160.22 |
| OOV | 2.30% | 2.69% | 2.59% | 1.82% | 2.84% |

**Table 5**. Specification of acoustic model

| Corpus | Corpus of Spontaneous Japanese (CSJ) 60 hours |
|---|---|
| Features | MFCC(12), ΔMFCC(12), ΔEnergy |
| #Phones | 43 |
| #States | 3,000 |
| Codebook size | 129 |
| #Mix. components | 128 |

### 3.3. Speech Recognition Results

We made ASR experiments using these models. Our decoder Julius 3.3[7] is used and sequential decoding is performed 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 that of supervised adaptation (58.9%). The result demonstrates that speaker adaptation based on the unsupervised speaker indexing improves the ASR accuracy.

The recognition performance for discussions is lower than that for lectures[6], since acoustic and language models are not completely matched to the discussions, while models for lectures are trained with the lecture speech corpus (CSJ).

## 4. CONCLUSION

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 reduced by model selection and normalized to suppress acoustic variations adaptively to the given input speech. These vectors are clustered and speaker models are trained for final indexing.

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

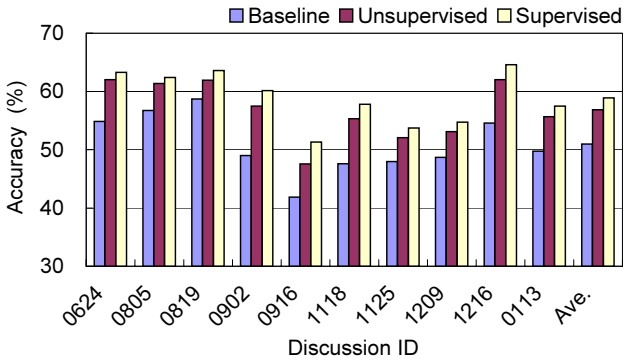


**Fig. 2**. Speech recognition result (word accuracy)

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 two models representing different linguistic features. With these models, we achieved word accuracy of 57%. 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.

## 5. REFERENCES

[1] 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.

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

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

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

[5] 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.

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

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