Speaker Model Quantization for Unsupervised Speaker Indexing

Soonil Kwon, Shrikanth Narayanan

Department of Electrical Engineering
Speech Analysis and Interpretation Lab
USC Viterbi School of Engineering
University of Southern California, U.S.A.
soonilkw@usc.edu, shri@sipi.usc.edu

Abstract

Speaker indexing sequentially detects points where speaker identity changes in a multi-speaker audio stream, and classifies each detected segment according to the speaker's identity. In unsupervised speaker indexing scenarios, there is no prior information/data about the speakers in the target data. To address this issue, a predetermined generic "speaker-independent" model set, called Sample Speaker Models (SSM), was previously proposed. While this set can be useful for more accurate speaker modeling and clustering without any target speaker models, an optimal method for sampling the models from such a set is still required. To address this problem, the Speaker Model Quantization (SMQ) method, motivated by Tree Structured Vector Quantization, is proposed. Experiments were performed with telephone conversations and broadcast news. Results showed that our new sampling approach outperformed the baseline by 5.5% absolute (37.7% relative) in error rate on 2 speaker telephone conversations, 10.7% absolute (42.5% relative) on broadcast news.

1. Introduction

Speaker recognition technology holds significant potential for enhancing our lives. One of the key speaker recognition applications is speaker indexing, the process of determining *who* is talking *when*. It is an integral element of rich speech data transcription and content-based data mining applications [1].

For automatic speaker indexing, ideally, we need information about the target speakers such as the number of speakers and the appropriate speaker models. However, in some scenarios, it is not easy to obtain a priori information about the target speakers. Consider for example speaker indexing applied to live broadcast news interviews. It may not be easy to obtain information about the reporters and interviewees in advance. Hence, unsupervised speaker indexing may be required. Assuming one is using streaming audio, we are limited to making any indexing decision with only current and previously seen speech data from the session. Furthermore, since the models of speakers are not available a priori for indexing, we need to create and update them on the fly. This leads to a number of challenges. In general, under these circumstances of sequential learning, data are not sufficient to build an adequate speaker model initially. Although a model can be roughly built, it is apt to cause decision errors due to potential uncertainty in the unsupervised learning.

To address the problem, we recently proposed a new method for creating and evaluating generic models, referred to as the Sample Speaker Models (SSM) [1]. This was built on the hypothesis that a speech data corpus, *independent* from the target data, can help initialize a model set for unsupervised speaker indexing. The Sample Speaker Models approach was found to provide better characteristics than other model bootstrapping methods used for unsupervised speaker indexing. In the original proposal for SSM, samples can be randomly picked from a pool of generic speaker models. However, the random selection method could not give optimally distributed sample models in the feature space. To select sample models optimally, we propose a novel method called Speaker Model Quantization (SMQ). Through the Vector Quantization process, speaker models in a pool can be quantized (categorized) to obtain some (optimal) representatives for speaker indexing. Our experiments showed that the SSM with SMQ method outperformed both SSM in conjunction with random selection and the baseline, Universal Background Model (UBM).

This paper is organized as follows: section 2 explains an unsupervised speaker indexing system and the use of generic models for bootstrapping; section 3 proposes the new Speaker Model Quantization (SMQ) method; section 4 describes our experiments and results; conclusion and future plan are described in section 5.

2. Unsupervised Speaker Indexing and Generic Models for Bootstrapping

2.1. Unsupervised Speaker Indexing

The block diagram of an unsupervised speaker indexing process is shown in Fig.1. The first step is front-end analysis that classifies audio samples into speech and different background audio (noise) types. Only the speech data are used

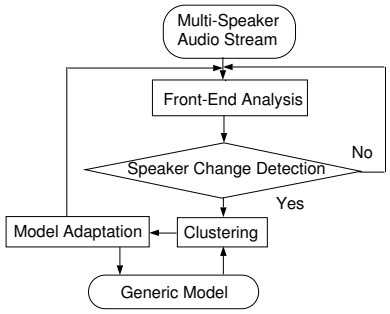

Figure 1: Block diagram of the unsupervised speaker indexing process with generic models.

in the next step, speaker change detection. In this step, the system sequentially detects whether a speaker changes in the middle of a speech analysis frame. In unsupervised scenarios, this step is executed without any knowledge about the identity or the number of speakers. For this detection, we use a Localized Search Algorithm (LSA) [2]. The analysis frame is divided into two equal length segments, which are then compared using the Generalized Likelihood Ratio (GLR) Test.

Speaker change detection step is important for the next step, speaker clustering. While we need relatively shorter analysis segments to detect the exact speaker changing points, longer segments are more helpful to classify speakers more accurately. Through the speaker change detection step, we can group relevant consecutive analysis segments into speaker-specific data segments. Those data segments are classified in the speaker clustering step. This step is executed using the Maximum Likelihood (ML) method. This approach usually assumes the availability of target speaker models. If a model is available, the just segmented data is used for further adaptation. If a model is unavailable, as is the case during the start of an unsupervised indexing scheme, appropriate model initialization is required (next section discusses details of model bootstrapping approaches). Next, audio samples after the boundary of the current speaker come into system, and the system repeats the previous steps until all data are exhausted.

2.2. Generic Models for Bootstrapping

Model initialization is critical in unsupervised speaker indexing. To build effective speaker models, enough training data are required. In the unsupervised scenario, there is no prior knowledge about the speakers. Further, only the data seen thus far can be used for modeling due to the characteristics of the sequential process. Such models that are roughly built can cause severe clustering errors. For alleviating the model initialization problem, generic models can be an alternative method. We build generic models of speakers, independent of the test (target) speakers, with the hypothesis that some speakers of the generic model set are acoustically close to the test (target) speakers. With this assumption, the initial generic model is built through training with data not directly related to the test condition. This can make it possible for an unsupervised system to run without training of the true (target) models.

There are a number of methods proposed for creating generic models such as Universal Background Model (UBM) and Sample Speaker Models (SSM). For example, suppose there are M male speakers and N female speakers in the generic speaker data pool. UBM is built pooling the entire speaker ($M + N$) data. SSM is a generic model set that we proposed previously [1]. We usually have a large number of speakers in a generic data pool. If all of speaker models in the pool are used for indexing, indexing errors may increase due to many acoustically similar models. Hence, an open problem is to choose the optimal number of speaker models that will balance between the false acceptance and false rejection errors. Also we need to pick a number, S, for the size of the model set which is smaller than the total number of speakers in the pool ($N_{target} < S \leq N_{pool}$). One possible approach to select the models is to randomly sample the speaker pool such as through the Markov Chain Monte Carlo (MCMC) method. Results in [1] showed that such an approach provided better performance than commonly used universal background or gender models.

While UBM involves "averaging" across a number of speakers, SSM does not. Each model of SSM is built with a single speaker data. For that reason, SSM has a smaller variance than UBM. In other words, SSM can be adapted faster and reflects a single speaker information better.

3. Speaker Model Quantization

The Sample Speaker Models (SSM) method is very useful in speaker indexing that operates without any prior knowledge of speakers. However, one critical issue with this SSM approach relates to finding the optimal number of sample models, and positions in the feature space, to use. A more principled approach is required in organizing the space spanned by the (generic) speakers for SSM. We describe here speaker model quantization for optimal speaker (model) sampling. This is motivated by Tree Structured Vector Quantization to obtain sample speaker models with the optimal number and positions in the feature space.

Some early attempts have been made to apply Vector Quantization (VQ) ideas to speaker recognition. Kinnunen et al. simply applied the vector quantization method to speaker identification [3]. They generated the VQ codebook and identified a target speaker with the Euclidean distance between target speaker data and the codebook. Pelecanos proposed a method of using Vector Quantization to generate a fast and robust approximation of the Gaussian Mixture Model [4]. Recently, Nishida and Kawahara proposed a novel method for model selection for speaker indexing. While GMM is a better method than VQ in terms of speaker recognition performance, it requires considerably more training data. When only small amounts of data are available

for training, VQ outperforms GMM. For that reason, a flexible framework was investigated: an optimal speaker model (GMM or VQ) was selected based on Bayesian Information Criterion (BIC) [5]. While, in all these previous efforts, VQ was used to replace Gaussian Mixture Models (GMM) or compensate for them in training, we utilize VQ for categorizing speakers based on their acoustical characteristics.

As mentioned earlier, in SSM, samples can be randomly picked from a pool of generic speaker models. However, the random selection method could not give optimal samples in the feature space. To obtain optimal samples, we propose a novel method called "Speaker Model Quantization (SMQ)". We do not quantize feature vectors but quantize speaker models in the feature space. In other words, the Speaker Model Quantization method is to select speaker models (GMM) that can represent several acoustically similar speakers well for unsupervised speaker indexing [Fig. 2]. The basic concept of SMQ originates from Tree Structured Vector Quantization (TSVQ) [6]. In our SMQ, a binary tree structure is used: each node splits two new nodes in each level. The Kullback-Leibler (KL) distance is used as the distortion measure. We decide to split or stop according to the decreasing rate of averaged distortion at each node [Fig. 3]:

$$R_i = \frac{D_i - D_{i+1}}{D_i} \gtrless \epsilon \qquad (1)$$

where D_i is the averaged KL distance of the i-th level.

To implement TSVQ for SMQ, a codebook can be built as follows:

- Step 1: Partition the root node (level 1) of the tree into 2 new subsets using the Lloyd algorithm.
- Step 2: Check the variation of the averaged distortion, R_1.
- Step 3: If R_1 is smaller than ϵ, stop splitting the nodes; otherwise, split the nodes.
- Step 4: Repeat node splitting and distortion checking (i=2,3,...) (from Step 1 to Step 3) until no nodes are left to split.

Based on this concept, it may possible to reduce the number of speaker models to a small and finite number. A speaker model giving the minimum distortion in each final node is chosen as a quantized speaker model representing the other models (acoustically similar speaker models) in the node.

The quantized speaker models are supposed to lie in the feature space with less overlap than in the case before quantization. There are two types of errors that are possible in the unsupervised speaker indexing. Type-1 is when a single target speaker is recognized as multiple speakers while type-2 is when multiple target speakers are indexed to a single speaker. Our previous experiments showed that type-1 errors are dominant with a larger number of sample speaker models. On the other hand, with a smaller number of sample speaker models, type-2 errors are dominant. It is an open problem

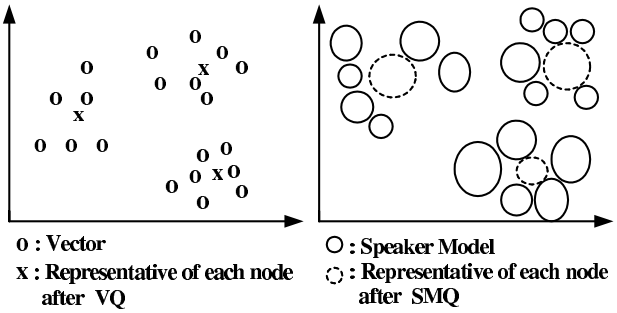

Figure 2: Vector Quantization vs. Speaker Model Quantization: (a) Vector Quantization, (b) Speaker Model Quantization.

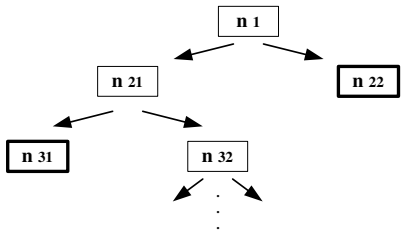

Figure 3: Example of a tree for Tree Structured Vector Quantization (TSVQ).

to choose the optimal number of speaker models while minimizing both these errors. After the SMQ process, only one speaker model exists in place of potentially several acoustically similar models, which reduces type-1 errors. However, it is not clear that SMQ can reduce type-2 errors. While the optimal number of samples is unknown, SMQ can give suboptimal number of sample speaker models from the given speaker model pool. In the next section, we describe the experimental results.

4. Experiments and Results

In this experiment, we compared three methods: Sample Speaker Models (SSM) with the Speaker Model Quantization (SMQ) method, Sample Speaker Models (SSM) with the random selection using Markov Chain Monte Carlo (MCMC) method, and the Universal Background Model (UBM). We had 400 speaker data (240 females and 160 males) in our speaker pool which come from the Speaker Recognition Benchmark NIST Speech (1999) corpus. For generic models, we used these telephone quality speech data. It was not easy to get data from hundreds of speakers from broadcast news audio data for training. Hence, with generic models trained with telephone speech, we indexed both telephone conversations and broadcast news. We used about 70 minutes of data for tests: 24 minute two speaker telephone conversations from the Speaker Recognition Benchmark NIST Speech (1999), and about 45 minute audio data from the HUB-4 Broadcast News Evaluation English Test

Table 1: *Error Rates of Unsupervised Speaker Indexing: Sample Speaker Models with Speaker Model Quantization (SMQ), Sample Speaker Models with MCMC, and Universal Background Model. Note that the figures in parentheses show the relative improvement to the baseline.*

Test set	UBM (baseline)	SSM	
		MCMC	SMQ
Telephone conversation	14.6% (-)	11.7% (19.9%)	9.1% (37.7%)
Broadcast news	25.2% (-)	18.1% (28.2%)	14.5% (42.5%)

Material (1999).

Before speaker indexing, we quantized 400 speakers in the pool using our Speaker Model Quantization (SMQ) method. As the result of SMQ, we got 70 speaker models representing our feature space. We also got another set of 70 speaker models using MCMC. For the random selection method (MCMC), we repeated the experiment three times to get the averaged result. A Universal Background Model (UBM) was built with the 400 speaker data in the pool.

Results are summarized in Table 1. The results of these experiments show that the case of using Sample Speaker Models (SSM) in conjunction with the Speaker Model Quantization (SMQ) method yielded the best performance. Table 1 shows that SSM-SMQ method outperformed SSM-MCMC by 2.6% absolute (22.2% relative) in the two speaker conversation case. It also shows a lower error rate than UBM by 5.5% absolute (37.7% relative) in this two speaker conversation case. With broadcast news, SSM-SMQ gave lower error rates than SSM-MCMC by 3.6% absolute (19.9% relative), and also much lower than UBM by 10.7% absolute (42.5% relative).

In sum, the generic model approach is helpful for unsupervised speaker indexing. However, approaches based on UBM are not adequate since they do not reflect speaker specific information well as they are constructed with data pooled from several different speakers. The experiment showed that the Sample Speaker Models (SSM) approach improved the unsupervised speaker indexing system over UBM. For SSM, we adopted the Markov Chain Monte Carlo (MCMC) method to pick the samples from the pool. However, this method could not get some (sub)optimal positions of speaker models in the feature space. For improved speaker model sampling, Speaker Model Quantization (SMQ) was proposed, and experimental results showed higher accuracy than SSM sampled by MCMC. These results indicate that SMQ gives better distributed samples in the feature space.

5. Conclusion

We presented a novel method for enabling unsupervised speaker indexing. For an unsupervised speaker indexing scenario, the Sample Speaker Models (SSM) approach helps to overcome some of the difficulties arising due to the lack of data for building true target speaker models. A key issue that was considered here relates to effectively sampling the SSM set. For a given feature space, some of the models can be severely overlapped, and some are farther apart, even if this formation can be thought to be inherently natural. We used a novel sampling method (SMQ) arriving at the optimally distributed speaker models. SMQ attained a measure of success to get some (suboptimal) positions of speaker models in the feature space. This method gave better experimental results than any other generic model considered thus far, including SSM with MCMC.

There are a couple of challenges that need further investigation in this context. The first relates to formally proving the optimality of SSM with SMQ through the use of information theoretic methods, such as the relative entropy between initial generic models and target models. The other question relates to the optimal size of the sampled speaker model set. This relates to the question of the "capacity" of speaker indexing i.e., model set size versus indexing performance trade offs. We need to find out what number of sampled models is optimal with the given speaker indexing condition. This is an open question that needs further research.

6. References

[1] Kwon, S. and Narayanan, S., "A Study of Generic Models for Unsupervised On-Line Speaker Indexing", Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop 2003, p.423-428.

[2] Kwon, S. and Narayanan, S., "A Method for On-Line Speaker Indexing Using Generic Reference Models", Proceedings of Eurospeech 2003, p.2653-2656, 2003.

[3] Kinnunen, T., Kilpeläinen, T., and Fränti, P., "Comparison of Clustering Algorithms in Speaker Identification", Proceedings of IASTED International Conference of Signal Processing and Communications (SPC 2000), p.222-227, 2000.

[4] Pelecanos, J., Myers, S., Sridharan, S., and Chandran, V., "Vector Quantization Based Gaussian Modeling for Speaker Verification", Proceedings of International Conference on Pattern Recognition, vol. 1, p.294-297, 2000.

[5] Nishida, M. and Kawahara, T., "Unsupervised Speaker Indexing Using Speaker Model Selection Based on Bayesian Information Criterion", Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, p.172-175, 2003.

[6] Gray, R. M. and Neuhoff, D. L., "Quantization", IEEE Transaction on Information Theory, Vol. 44, p.2325-2383, No. 6, 1998.