A STUDY OF GENERIC MODELS FOR UNSUPERVISED ON-LINE SPEAKER INDEXING

Soonil Kwon, Shrikanth Narayanan

Department of Electrical Engineering, Speech Analysis and Interpretation Lab,
and Integrated Media Systems Center
University of Southern California, CA, U.S.A.
soonilkw@usc.edu, shri@sipi.usc.edu, http://sail.usc.edu

ABSTRACT

On-line speaker indexing sequentially detects the points where a speaker identity changes in a multi-speaker audio stream, and classifies each speaker segment. This paper addresses two challenges: The first relates to monitoring which requires on-line processing. The second relates to the fact that the number/identity of the speakers is unknown. The indexing needs to be made in a unsupervised process. To address these issues, we apply a predetermined generic speaker-independent model set, Sample Speaker Model(SSM). This set can be useful for more accurate speaker modeling and clustering without requiring training models on target speaker data. Once a speaker-independent model is selected from the sample models, it is adapted into a speaker-dependent model progressively. Experiments were performed with Speaker Recognition Benchmark NIST Speech(1999). Results showed that our new technique, simulated using Markov Chain Monte Carlo Method, gave 92.47% indexing accuracy on telephone conversation data.

1. INTRODUCTION

Speaker indexing, the process of determining who is talking when, is an integral element of speech data monitoring and mining applications. Consider, for example, applications such as meeting/teleconference monitoring, archiving and browsing. However, it is impossible to attend all relevant meetings face to face. Multimedia meeting or teleconference monitors and browsers can be useful for getting meeting information, such as who is saying what and when, remotely through the on-line or off-line systems [1] [2].

These applications commonly include a speaker indexing process that tags speaker-specific portions of data to pin point who is talking when [3]. Off-line speaker indexing can be used for record keeping, but it is not appropriate for real-time meeting or teleconferencing systems. Recently we proposed an on-line method that picks out the speech segments from an audio stream and classifies them by speakers [12].

On-line speaker indexing method can only be sequentially executed. In other words, assuming streaming audio, we are limited in making any indexing decision with only current and previously seen speech data in the session. Moreover the indexing problem gets more difficult if there is no prior knowledge about the target speakers in the data including the number of speakers. Since the models of speakers are not available a priori for indexing, we need to build and update them on the fly. This implies a number of challenges. In general, under these circumstances, data are not sufficient to build a speaker model. Although a model can be roughly built, it is apt to cause decision errors. To address the problem, we need a generic model to enable effective bootstrapping.

There are two kinds of generic models that have been proposed previously: Universal Background Model(UBM) and Universal Gender Model(UGM). In this paper, we propose a new method for creating and evaluating generic models, referred to as the Sample Speaker Model(SSM). This is built on the hypothesis that an independent speech data corpus can help initialize a model set for unsupervised indexing. Samples are picked from a pool of generic speaker models using Markov Chain Monte Carlo(MCMC) method. The sample model set is predetermined by training. Note that the speakers in the training data are independent of the testing data. In other words, the generic model set can be used for initializing/bootstrapping any speaker indexing process. This model set can be referred during speaker clustering with the test data. After clustering, a selected model is continually adapted with the test data that are used for clustering [Fig.1].

For clustering, the size of analysis frame is fixed. Large frame size helps toward a correct indexing decision, since long frame includes enough speaker information for indexing. However, it is apt to miss speaker changes. To solve this problem, a small indexing frame size is used in conjunction with a robust speaker change detection process to improve the precision [12]. We use the Generalized Likelihood Ratio(GLR) Test for speaker change detection [3]. Though the GLR Test can be unstable for small amounts of

0-7803-7980-2/03/$20.00 © 2003 IEEE

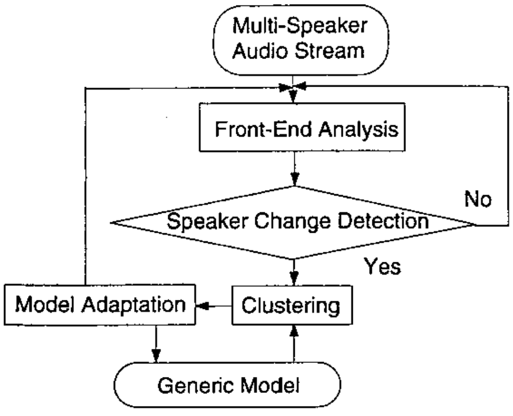

Fig. 1. *Block diagram of on-line speaker indexing.*

analysis data, clustering can help compensate for this instability.

We experimented with the "Speaker Recognition Benchmark NIST Speech Corpus(1999)". The experimental results show that our on-line speaker indexing can achieve a recognition rate comparable with a state of the art off-line system [3].

This paper is organized as follows: section 2 explains our on-line speaker indexing system; section 3 describes generic models for on-line speaker indexing; section 4 is about clustering and model adaptation; section 5 and section 6 describe our experiments and results respectively; our conclusion and future plan are described in section 7.

2. ON-LINE SPEAKER INDEXING

The block diagram of the on-line 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 for the next step, speaker change detection. In this step, the system sequentially detects whether a speaker changes in the middle of speech analysis frame without any knowledge about the identity or the number of speakers. For this detection, we use a Localized Search Algorithm(LSA) [12]. Two segments within the window are compared using the Generalized Likelihood Ratio(GLR) Test. Suppose there are two feature vector sets, X_1 and X_2, coming from each segment [3] [9]. Each segment contains speech from a single speaker. Hypothesis, H_0, is that the speakers in two segments are same, and H_1 is that the speakers are different. Let $L(X_1;\lambda_1)$ and $L(X_2;\lambda_2)$ be the likelihood of X_1 and X_2 where λ_1 and λ_2 represent model parameters that maximize each likelihood. Similarly let X be the union of X_1 and X_2. $L(X;\lambda_{1+2})$ is the maximum likelihood estimate for X. Then

$$GLR = \frac{L(X;\lambda_{1+2})}{L(X_1;\lambda_1)L(X_2;\lambda_2)} \qquad (1)$$

We apply thresholding on GLR to determine the latent changing point. Speaker change detection step is important for the next step, speaker clustering. We can get speech data that is longer than 2 seconds. It is very helpful for better speaker clustering, because more speech data usually help in representing a speaker better. However this step cannot be ensured to be perfect. If we falsely detect a speaker changing point, we can compensate for the error through the speaker clustering step. However, if we skip the real changing point, the clustering step cannot recover it. For these reasons, we tightly detect changing points to avoid the skip, although some points can be wrongfully detected as changing points. After clustering, a generic model for the current speaker is adapted into the current speaker dependent model. The adapted model is replaced with the original model before adaptation or inserted in the generic model set. 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.

There have prior efforts that have been reported on on-line speaker segmentation and clustering [6] [7]. Online speaker processing means that speaker recognition and model construction can be performed sequentially without storing all the testing data in advance. Furthermore, in many scenarios, there is no prior knowledge about the identity or the number of speakers involved. Therefore, for such on-line speaker indexing applications, it is difficult to build true speaker models in advance. Without the good initial models for speaker indexing, we cannot effectively build/update speaker models sequentially and incrementally. Note that that sequentially constructed models can not represent speakers well due to small initial amount of data. Since the training is unsupervised, this problem also potentially leads to continual error propagation. We try to solve this critical drawback using an alternative method.

3. GENERIC MODELS FOR BOOTSTRAPPING

To build effective speaker models, enough training data are required. When the on-line process starts, there is no prior knowledge about the speakers. Only the data seen thus far can be used for modeling due to the characteristics of the online process. Such models that are roughly built can cause severe clustering errors. The question then is if we can find a method for alleviating the model initialization problem: How can we build speaker models for on-line speaker indexing without prior knowledge/data about target speakers? Generic models can be an alternative method. We build generic models of speakers that are independent of the test set speakers with the hypothesis that some speakers of the reference set are acoustically close to the test speaker [11]. Although we do not know the exact number of speakers, we assume that the number is finite. 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 on-line system to run without training of the true models.

There are at least three possibilities that one can consider for creating generic models: Universal Background Model(UBM), Universal Gender Model(UGM), Sample Speaker Model(SSM). For example, suppose there are *M* male speakers and *N* female speakers in the generic speaker data pool. The UBM is built pooling the entire speaker (*M* + *N*) data. UGM includes two models: One is for male speakers, and the other is for female speakers. The male UGM is trained with *M* male speaker data, the female UGM is trained with *N* female data. SSM is a new generic model set that we propose in this paper. At first, we pick *S* speakers, the number of which is smaller than the total number of speakers in the pool. Then *S* speaker models are trained. Markov Chain Monte Carlo(MCMC) method can be used for sampling [8]. The initial distribution of random number is from the Universal Background Model(UBM) that represents all the speakers in the pool. While UBM and UGM involve "averaging" across a number of speakers, SSM does not.

4. CLUSTERING AND MODEL ADAPTATION

The segments obtained from the speaker change detection are indexed in terms of speakers, and then the corresponding models are adapted with the new data. For clustering, we use speaker models from our predetermined generic model: UBM, UGM, SSM. In UBM and UGM cases, the generic pooled model is adapted to create speaker dependent models. When SSM is used, the "Speaker independent" models are adapted into speaker dependent models. The likelihood of a speaker segment is calculated with sample speaker models, and the model with maximum likelihood is selected and adapted. Model adaptation is executed by Maximum a Posteriori(MAP) scheme. As the amount of data increases towards infinity, the MAP estimate converges to the ML estimate [10]. The MAP adaptation on a Gaussian Mixture Model(GMM) is straightforward [14]. Given the adaptation vectors $X = \{x_1, x_2, ..., x_T\}$, we compute the probability, $Pr(i|x_t)$:

$$Pr(i|x_t) = \frac{w_i p_i(x_t)}{\sum_{l=1}^{M} w_l p_l(x_t)} \tag{2}$$

where w_i is the weight of each mixture of GMM, and p_i is the probability of input, x_t, in each mixture. M is the number of mixtures. In this system, means, $\hat{\mu}$, and weights, $\hat{w}$, of GMM are updated as:

$$\hat{\mu_i} = \alpha_i^m E_i(x) + (1 - \alpha_i^m)\mu_i \tag{3}$$

$$\hat{w}_i = [\alpha_i^p n_i / T + (1 - \alpha_i^p) w_i]\gamma \tag{4}$$

where γ is a scale factor. α_i^m and α_i^p are data-dependent adaptation coefficients which are defined as:

$$\alpha_i^p = \frac{n_i}{n_i + r_\rho} \tag{5}$$

where r_ρ is the fixed relevance factor. n_i are the sufficient statistics of mixtures, and $E_i(x)$ are the re-estimation of mixtures which are defined as:

$$n_i = \sum_{t=1}^{T} Pr(i|x_t) \tag{6}$$

$$E_i(x) = \frac{1}{n_i} \sum_{t=1}^{T} Pr(i|x_t) x_t \tag{7}$$

We assume that speaker models in the reference set are independent of the testing speech data. From this assumption, we can expect that the initial adaptation/learning rate to the "true" speaker data should be rapid.

5. EXPERIMENTS

The Speaker Recognition Benchmark NIST Speech(1999) is used for the experiments reported in this paper. For the generic model, we used 100 speakers (50 male speakers and 50 female speakers) who were randomly selected from the training data in Speaker Recognition Benchmark NIST Speech(1999). Since this corpus consists of telephone conversations, there was no significant background noise to adversely affect the recognition. For each speaker model training, about one minute of speech data were used. We tested our on-line speaker indexing system also using independent portions of this data corpora. Testing was executed with about 24 minute audio data from the Speaker Recognition Benchmark NIST Speech(1999). The length of each speech audio sequence was about one minute. Every sequence included two speaker telephone conversations. One third of sequences included mixed gender(one male and one female) conversations. The other sequences include two males or two female speakers. No speaker used for building the generic model participated in the test conversations.

We report two experiments in this paper. One of them was relevant to the convergence of model adaptation. For better speaker change detection, we need to use short speech segments for speaker clustering. However, we don't know what length of segment is optimal. Also it is not easy to test the convergence properties of model adaptation. To investigate both of these questions experimentally, we tested speaker identification accuracy with various lengths of data segments and with the three kinds of generic models. The other experiment concerned exploring the performance of the generic models. We tested which of the three generic models(UBM, UGM, SSM) showed the best on-line speaker indexing performance on two speaker telephone conversations. In the Sample Speaker Model case, we applied various number of samples (i.e., 4,8,16,32,64 sample models).

Since long silences have an adverse effect on speaker recognition, we eliminated silence which is longer than 100 msec and lower than -40 dB. Experimental data were sampled at 8000Hz. As feature vectors, we used 26 channel, 24 dimensional Mel Cepstrum vectors. We also used 30 msec Hamming window that was shifted by 10 msec. Speaker models were Gaussian Mixture Models(GMM) with 16 mixtures.

6. RESULTS

Sequential speech data extracted from the input audio stream were chopped into segments by speakers, and the segments were classified through the generic model. The result of the first experiment is shown in Table 1. In this experiment, we looked into the property of adaptation and convergence for on-line speaker indexing under various conditions. We randomly picked 32 speakers from the 100 speaker pool. The Universal Background Model(UBM) and Universal Gender Model(UGM) were built using data from those 32 speakers. Multiple speakers in Table 1 imply that multiple speakers participated in conversations. For example, suppose that we had 4 speakers in a test. We randomly picked another 32 speakers from the pool, and chose 4 speakers for testing. Each speaker had about 1 minute speech data, so totally about 4 minutes of data were used for on-line processing. The current speaker was changed at every minute interval. The point is that the indexing process was executed only sequentially without any prior speaker models. While the first minute segments passed through the indexing system, some speaker changing points might be detected as changing points that could be false changing points. The first adapted speaker model and the generic model(s) were compared with the group of segments that were hypothesized as speech of a certain speaker. After one speaker model was adapted from the generic model, two models were compared. Whenever a speaker was changed, the system looked for the next speaker model. As speakers changed, the number of models we had to compare increases which in turn might affect the error rates. In other words, as the number of speaker candidates in a conversation increased, the accuracy is expected to drop.

Table 1. *Length of Segment, Number of Speakers, and Generic Models vs. Accuracy*

Length of Segment(sec)	Generic Model	Number of Speakers		
		1	2	4
1	SSM	97.8%	100%	77.2%
	UGM	96.1%	95.7%	93.4%
	UBM	97%	97%	64.7%
2	SSM	98.4%	100%	100%
	UGM	98.1%	100%	99.1%
	UBM	98%	100%	99.1%
4	SSM	99.5%	100%	100%
	UGM	99.5%	100%	98.2%
	UBM	99.6%	100%	100%
8	SSM	100%	100%	100%
	UGM	99%	100%	96.4%
	UBM	100%	100%	100%

When the length of the segments was increased (e.g., to 8 seconds), speaker indexing accuracy was almost 100% in most cases. It means that 8 seconds of speech data include enough information to represent a speaker in these experiments. However, what are the implications of using shorter data segments for speaker detection? For example, consider a 20 second two speaker conversation, and we used 8 second long analysis segments. Suppose that the one of the speaker's speech lies between 5 and 7 second (measured from the beginning). Although the first 8 second analysis segment included two speakers, it was recognized as one speaker. To detect shorter speech episodes, we should use as short an analysis segments as we can. However, shorter (e.g., 1 second) segments could not capture a speaker's information adequately in our experiment. Empirically, we determined from this first experiment, that a 2 second analysis segment was optimal.

In Table 1, we used three kinds of generic models: Sample speaker Model(SSM), Universal Gender Model(UGM) and Universal Background Model(UBM). SSM had the most stable performance whatever the number of speaker candidates and the length of segment are. Higher accuracy also meant model adaptation was better. Whenever a segment assigned to a speaker, the speaker model was adapted. Therefore good performance meant that it was adapted and converged well.

From the results of the first experiment, we used 2 second analysis segments for the next experiment. Fig.2 shows which of the three generic models was the best in the on-line speaker indexing of two speaker telephone conversations. Initial Universal Background Model(UBM) was a unitary Gaussian Mixture Model that was trained from 100 speaker speech data in the pool. Universal Gender Model(UGM) had two models: male and female. The Sample Speaker Model(SSM) set can have varied number of models. In our experiments, we used 4, 8, 16, 32, and 64 model sets to find

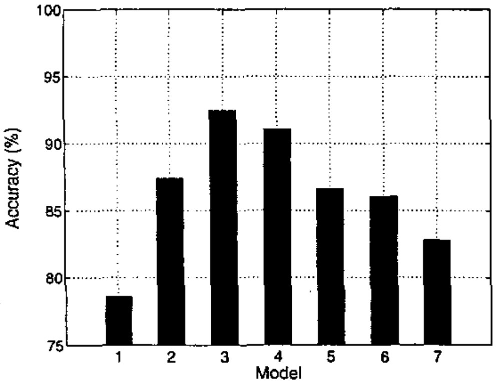

Fig. 2. *Generic Models vs. Accuracy : 1. SSM (4 Samples), 2. SSM (8 Samples), 3. SSM (16 Samples), 4. SSM (32 Samples), 5. SSM (64 Samples), 6. UGM , 7. UBM. Note that 'samples' here refer to those drawn from the generic model pool by MCMC.*

empirically the optimal number of samples for on-line two speaker conversation indexing. When the number of speakers was smaller than 16, the accuracy was below 90%. The reason is that 4 and 8 models were not enough to recognize two speakers, as they could not have adequate discriminatory power in our feature space. While the 32 and 64 model cases performed better than the 4 model case, they were worse than the 16 model set. As the number of models increased, too many similar models occupied the feature space. In this situation, one (test) speaker could be recognized as two or more (model) speakers. From this experiment, 16 was found to be the optimal number of sample speaker models.

The results of Universal Background Model(UBM) and Universal Gender Model(UGM) cases were worse than for the 16-SSM. UBM was built with 100 speaker data, and UGM consisted of a male model and a female model that were built with 50 male and 50 female speaker data, respectively. Each model of SSM was a certain generic speaker model. For that reason, UBM and UGM had larger variances, and each speaker model was adapted from those initial models. Although the model variance could be adapted as well, it is difficult to represent a speaker well with small amounts of data. UGM is better than UBM because gender models surely had relatively smaller variances. These experiments showed that the automatic selection of initial models in a way most similar to the final target models leads to better and faster model adaptation and convergence.

The condition for the off-line indexing case is different from that for the on-line indexing case since the former case can be processed iteratively but the latter is not. We applied the 16-SSM to the off-line system to see whether it is useful in the off-line environment or not. The result showed that the speaker indexing was converged at the first iteration. It may imply two possibilities: The first possibility is that the off-line unsupervised speaker indexing with 16-SSM has the same performance as that of the on-line system. The other is that the selected initial models are adapted and converged well in the on-line unsupervised speaker indexing with 16-SSM.

7. CONCLUSION

We presented a novel method for enabling on-line speaker indexing. For an on-line process without any prior knowledge of the speakers, a generic model set was inserted into the general speaker indexing algorithm. This generic model helped the on-line speaker indexing system to overcome the difficulty due to the lack of data for building target speaker models. In particular, the Sample Speaker Model(SSM) approach showed better performance than the other generic model methods. With SSM, we do not have to get the same speaker training data for indexing in advance. One critical issue of this SSM approach is how to find the robust number of sample models to use. In this paper, we adopted the Markov Chain Monte Carlo(MCMC) method to pick the samples from the pool.

We used telephone conversation data to evaluate the performance of our algorithm. The condition that yielded the best performance in our experiments was using 2 second analysis segments in conjunction with 16 sample speaker models. The total error rate was 7.53%, so it reduced about 10% absolute compared with the UBM case(17.21%).

There are at least three key issues worth considering to further improve the overall performance of on-line speaker indexing: strategies for effectively sampling the SSM set, robustly detecting speaker changes, and adapting speaker models. Some of the models can be severely overlapped, and some are farther apart, even if this formation can be thought to be inherently natural. More experiments are required in organizing the generic speaker pool for SSM. Another important step, speaker change detection, was performed with GLR. This method is acceptable, but has room for further improvement using other statistical approaches. For model adaptation, we used MAP. This is good for controlling the speed of adaptation, but we will try to find better methods.

8. ACKNOWLEDGEMENTS

The authors would like to thank their colleagues in the USC Speech Analysis and Interpretation Lab(SAIL). This research was partially supported by NSF under cooperative agreement No. EEC-9529152 through the Integrated Media Systems Center and by the Department of Army under contract number DAAD 19-99-D-0046. Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the aforementioned funding agencies.

9. REFERENCES

[1] Kwon, S. and Narayanan, S., "Speaker Change Detection Using a New Weighted Distance Measure", International Conference on Spoken Language Processing, vol. 4, p.2537-2540, 2002.

[2] Yang, J., Zhu, X., Gross, R., Kominek, J., Pan, Y., and Waibel, A., "Multimodal People ID for a Multimedia Meeting Browser", Proceedings of 7th ACM International Conference on Multimedia, Part 1, p.159-168, 1999.

[3] Rosenberg, A., Gorin, A. , and Parthasarathy S., "Unsupervised Speaker Segmentation of Telephone Conversations", International Conference on Spoken Language Processing, vol. 1, p.565-568, 2002.

[4] Siu, M-H., Yu, G., and Gish, H., "An Unsupervised, Sequential Learning Algorithm for the Segmentation of Speech Waveforms with Multiple Speakers", Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, p.189-192, 1992.

[5] Lu, L., Zhang,H. -J., and Jiang, H., "Content Analysis for Audio Classification and Segmemtation", IEEE Trans. on Speech and Audio Processing, Vol. 10, p.504-516, No. 7, 2002.

[6] Wu, T., Lu, L., Chen, K., and Zhang, H., "UBM-Based Real-Time Speaker Segmentation for Broadcasting News", Acoustics, Speech, and Signal Processing Proceedings, IEEE International Conference on , vol. 2, p.193-196, 2003.

[7] Liu, D. and Kubala. F., "Online Speaker Clustering", Acoustics, Speech, and Signal Processing, Proceedings. IEEE International Conference on, vol 1, p.572-575, 2003.

[8] Davy, M., Doncarli, C., and Tourneret, J., "Supervised Classification Using MCMC Methods", ICASSP'2000, International Conference on Acoustics, Speech and Signal Processing, p.33-36, 2000.

[9] Solomonoff, A., Mielke, A., Schnidt, M., and Gish, H., "Clustering Speakers by Their Voices", Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, p.12-15, 1998.

[10] Woodland, P. C., "Speaker Adaptation: Techniques and Challenges", in Proc. IEEE Workshop Automatic Speech Recognition and Understanding, Keystone, Colorado, Dec., p.85-90, 1999.

[11] Wu, J. and Chang, E., "Cohorts Based Custom Models for Rapid Speaker and Dialect Adaptation", Eurospeech2001, p.1261-1264, 2001.

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

[13] Bonastre, J-F., Delacourt, C., Fredouille, T., Merlin, T., and Wellekens, C., "A Speaker Tracking System Based on Speaker Turn Detection for NIST Evaluation", Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, p.1177-1180, 2000.

[14] Liu, M., Chang, E., and Dai, B. -Q., "Hierarchical Gaussian Mixture Model for Speaker Verification", International Conference on Spoken Language Processing, vol. 2, p.1353-1356, 2002.