On-Line Speaker Indexing

1. Research Team

Project Leader: Prof. Shrikanth Narayanan, *Electrical Engineering*

Post Doc(s): Panayiotis Georgiou

Graduate Students: Soonil Kwon

2. Statement of Project Goals

Unsupervised speaker indexing sequentially detects points where a speaker identity changes in a multi-speaker audio stream, and categorizes each speaker segment, without any prior knowledge about the speakers. This project addresses two challenges: The first relates to sequential speaker change detection. The second relates to speaker modeling in light of the fact that the number/identity of the speakers is unknown. To address this issue, a predetermined generic speaker-independent model set, called the Sample Speaker Models (SSM), is proposed. This set can be useful for more accurate speaker modeling and clustering without requiring training models on target speaker data.

3. Project Role in Support of IMSC Strategic Plan

Speaker indexing, the process of determining who is talking when, is an integral element of speech data monitoring and content-based data mining applications. Consider, for example, applications such as meeting/teleconference monitoring, archiving and browsing. A key motivation arises from the fact that it is impossible or tedious to attend all relevant meetings face to face. Multimedia meeting or teleconference monitors and browsers can be useful for conveniently obtaining meeting information, such as who is saying what and when, remotely through on-line or off-line systems. Specially, these applications commonly include a speaker indexing process that tags speaker-specific portions of data to pin point who is talking when.

4. Discussion of Methodology Used

The first step is front-end analysis where the incoming audio samples are classified into foreground speech and other background audio (noise) types. Generally audio data can be categorized into four broad classes: speech, music, environmental noise, and silence. In speaker indexing, we only need speech/nonspeech discrimination. When there is background noise or music, it is likely to be overlapped with speech. Corrupted speech is not easily discriminated from noise. Since it is critical that we should not lose any speech data, the focus of the classification is to minimize false rejection, perhaps, even at the cost of false acceptance. Usually, for speech/nonspeech discrimination, a zero-crossing rate and short-time energy are used. It is well known that speech has a higher level of variation in the zero crossing rate. Only the speech data are used for the next step, speaker change detection. In this step, the system sequentially detects whether a speaker change occurs in the middle of a speech analysis frame,

without assuming any specific knowledge about speakers. Once the speaker change detection determines the boundary, all the data between the speaker change points are used for speaker clustering. In the clustering step, we use speaker models from a predetermined generic model set. After clustering, the speaker independent generic model is adapted into an appropriate speaker dependent model during the indexing process. The adapted model is replaced with the original model before adaptation or inserted back in the generic model set. When new audio samples after the boundary of the current speaker come into the system, the previous steps are repeated until all data are exhausted.

5. Short Description of Achievements in Previous Years

We presented a novel method for enabling unsupervised speaker indexing. For an unsupervised sequential process without any prior knowledge about the speakers, a generic model set was incorporated into the general speaker indexing framework. This generic model was shown to help the unsupervised speaker indexing system to overcome some of the difficulties arising due to the lack of data for building true target speaker models. In particular, the Sample Speaker Models (SSM) approach showed better and more stable performance than the other generic model methods such as Universal Background Model (UBM) and Universal Gender Models (UGM). Since these generic models do not contain the type of speakers training data in the initial for indexing, this implies that we do not have to retrain speaker models whenever we test with different speakers.

5a. Detail of Accomplishments During the Past Year

We used telephone conversation data and broadcast news 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 for 2 speaker conversations. The total error rate was 7.53%, which was about 10% lower absolutely compared with the UBM case (17.21% relative). In the case of 4 speaker conversations, 16 sample speaker models performed the best with an error rate, 10.4%, which is about 30% absolutely better than that of UBM. As the number of speakers present in conversations increases, the error rate of UBM increased at a much higher rate than that of 16-SSM model set. In the experiment with the broadcast news, the actual number of speakers was not known. The news clips considered had between two and six speakers. More samples were required than those in the 4 speaker telephone conversations to cover the wide range in the number of speakers. The result showed that 64 was the optimal number of speakers for broadcast news clips with an error rate of 12.8%, which is about 20% absolutely better than that of UBM (31.8%). The result shows that the performance of 64-SSM was stable: the error rate was only 20% in the worst case.

6. Other Relevant Work Being Conducted and How this Project is Different

Several efforts have been reported on speaker indexing. Methods based on speaker verification using speaker subspace for speaker indexing were proposed by Nishida and Ariki. In this paper, a speaker model was initialized and then the next speech segment was verified if it was from the same speaker as the first one. They used only 1 second segments, and these were too short to build an initial speaker model. Some segments including the speech of more than 2 speakers

could not be correctly clustered without the speaker change detection. Rosenberg et al. [5] used the Generalized Likelihood Ratio (GLR) Test for initial segmentation of speaker indexing. After initial segmentation, speaker models were constructed and then repeatedly segmented. Their process focused on the iterative segmentation and clustering that was only for the off-line speaker indexing systems. Solomonoff introduced the metric based on purity and completeness of clusters for speaker clustering. With this method, even though it is not necessary to train speaker models, it is not found to be robust to environmental noises. There are other efforts that have been reported on on-line speaker segmentation and clustering without prior knowledge of speakers and speaker models. The Universal Background Model (UBM) was used to classify feature vectors by Wu. Liu used the Hybrid speaker clustering method, which utilized both the dispersion and GLR threshold.

7. Plan for the Next Year

There are a couple of issues that need further investigation in this context. One critical issue with this SSM approach relates to finding the optimal number of sample models and positions in the feature space to use. 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. A more principled approach, with supporting experiments, is required in organizing the space spanned by the (generic) speakers for SSM, such as feature space or speaker quantization for optimal speaker (model) sampling. Lastly, higher level linguistic information and multi-modal features can be integrated to overcome the limitations of the speaker recognition based on just spectral envelope speech features.

8. Expected Milestones and Deliverables

Robust and elaborate on-line multi-speaker indexing algorithm.

9. Member Company Benefits

N/A

10. References

[1] Kwon, S. and Narayanan, S., ``Speaker Change Detection Using a New Weighted Distance Measure'', International Conference on Spoken Language Processing, Volume Two, p.2537-2540, 2002.
[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] 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.
[4] Nishida, M. and Ariki, Y., "Speaker Indexing for News Articles, Debates, and Drama in Broadcasted TV Programs", IEEE International Conference on Multimedia Computing and systems, Volume Two, p.466-471, 1999.
[5] 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.