Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 6, Issue, 03, pp.5717-5720, March, 2014

INTERNATIONAL JOURNAL
OF CURRENT RESEARCH

ISSN: 0975-833X

RESEARCH ARTICLE

A MULTIMODAL APPROACH TO SPEAKER DIARIZATION USING TABU SEARCH

*Sasikala, S. and Dr. Balamurugan, P. Ph.D

Department of Computer Science and Engineering, K.S.R. College of Engineering Tiruchengode, Namakkal, India

ARTICLE INFO

Article History:
Received 16th December, 2013
Received in revised form
19th January, 2014
Accepted 28th February, 2014
Published online 25th March, 2014

Key words:

Speaker diarization,
Support vector machines,
Tabu search,
Joint audiovisual processing,
Audio classification,
Clustering,
TV Talk shows.

ABSTRACT

In this paper proposes solutions for speaker diarization in TV talk shows for multimodal approaches. Both audio and video data can be taken for multimodal approach. In this paper can decomposes two levels such as, the reliable datasets can be formed for TV shows and SVM is used to classify the audio and video data based on unsupervised approach. Both audio and video data can be assembled by the association of visual and audio descriptors. Tabu search is introduced for improve the accuracy of the searching method. Once audio and visual features have been extracted, the system taking through collect the learning and classifying audiovisual frames based on SVM using Tabu search method. Time complexity will be reduced by using Tabu search. The result will produce the better output by using Tabu search algorithm. There are two schemes are measured for audio and video data such as audio-only classification scheme and parallel audio/visual classification scheme. The improvements of speaker diarization methods can be established effectively.

INTRODUCTION

Speaker diarization consists of segmenting and clustering a speech recording into speaker homogenous regions, so that given an audio track of a meeting the system will be categorize and label the different speakers automatically ("who spoke when?") applications for speaker diarization algorithms include speech and speaker indexing, document content structuring, speaker recognition (in the presence of multiple or competing speakers), to help in speech-to-text transcription, speech translation and, more generally, Rich Transcription (RT), a community within the current state-of-the-art technology has been developed. Speaker diarization has service in a majority of applications associated to audio and/or video document processing, such as information retrieval. In the speaker diarization process have two stages. First is the segmentation step, which locates segment boundaries based on acoustic changes in the signal. Second is the clustering step, which regroups segments coming from the same speaker into a clusters. The video samples can be extracted into number of frames; the frames can be taken as an input for processing. Face detection is used to identify the faces in the frames of the visual content. The lip activity can be recognized by using the features of the visual content. The audio samples can be determined into the histograms. SVMs (Support Vector Machines) are a useful technique for data classification. SVM is primarily a classier method that performs classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels. Each instance in the training set contains one target value (class label) and several attributes (features). The goal of a classifier is to produce a model able to predict target values of data instances in the testing set, for which only the attributes are known.SVM classification is used to classify and cluster the audio and video data. The clustered audio and video data can be assembled by using Tabu search method. It is one of the local algorithms used in SVM classification. Tabu search is used to arrange the audio and video samples in clustered class labels and predicts the optimal solutions.

Tabu search is a metaheuristic local search algorithm that can be used for solving combinatorial optimization problems. Local searches take a possible solution to a problem and check its immediate neighbors in the hope of finding an improved solution. Local search methods have a tendency to become stuck in suboptimal regions or on plateaus where many solutions are equally fit. Tabu search enhances the performance of these techniques by using memory structures that describe the visited solutions or user-provided sets of rules. If a potential solution has been previously visited within a certain short-term period or if it has violated a rule, it is marked as "Tabu" (forbidden) so that the algorithm does not consider that possibility repeatedly. Tabu search can be embedded with SVM classification for producing the accurate result.

***Corresponding author:** *Sasikala, S. Department of Computer Science and Engineering, K.S.R. College of Engineering Tiruchengode, Namakkal, India.*

RELATED WORK

In the paper, investigations are done on the EPAC corpus, mainly containing conversational documents. The application domains, from broadcast news, to lectures and meetings, vary greatly and pose different problems, such as having access to multiple microphones and multimodal information or overlapping speech. In this paper (Bigot *et al.*, 2010) based on spontaneous speech like conversational programs and it has huge amount of audio visual documents. Five different roles can be defined such as, Anchor, Punctual Journalists, Journalists, Punctual others and others. In this paper mostly considered on the basis of TV talk shows. This paper (Friedland *et al.*, 2009) says that the first audiovisual speaker diarization systems were proposed. On the basis of this surveillance, we propose completely novel multimodal speaker diarization architecture, well-adapted to talk-shows. In (Bendris *et al.*, 2010) shows that separating the talking faces and non-talking faces by detecting lip motion. In the previous work, (Vallet *et al.*, 2010) focused on robust visual features may prove very useful for the initialization of a top-down speaker diarization system and SVM classification can be considered and also can be extracted the images with clothing information.

In (Friedland *et al.*, 2009) Acoustic features can be combined with the compressed domain video features. In (Dielmann 2010) shows that a novel multimodal method for identifying unregistered speakers in a TV talk-show using a semi-supervised learning approach based on Support Vector Machines. In this paper, (Anguera *et al.*, 2012) deals with they reviewed the current state-of-the-art, focusing on research developed since 2006 that relates predominantly to speaker diarization for conference meetings. Finally, the system present an analysis of speaker diarization performance as reported through the NIST Rich Transcription evaluations on meeting data and identify important areas for future research.

The multimodal approach

Multimodal Approach based on speaker diarization system such as speaker recognition or speaker identification methods. It involves three different steps that are,
(i) Feature extraction,
ii) Collecting Training Examples, and
iii) Hypothesized speakers classification.

Feature extraction

In visual descriptors is used to consider features characterizing the clothing of the TV show-participants, building upon the method initially proposed. Indeed, though the field size of the shots are varying (from long shots to close-ups), most of the time the person talking is seen on-screen. However, the use of a facial recognition system is rather difficult in our task due to the camera movements, the changing field-size and angles of shot, the changing postures of the filmed persons and the varying lighting conditions. The approach that we choose has the advantage that features relating to on-screen persons' clothing can be extracted even more robustly than the persons' facial features. Another motivation is the fact that in the TV production domain, the clothing of the participants on a TV set is often carefully chosen.

Collecting Training examples

Shot detection

In the training set, the features can be given for identifying the faces for accurate results. In the shot detection, the human faces can be detected using speaker detection.

Lip activity detection

In the lip activity detection, the human faces can be extracted and the lip features is already given in the training set such as,

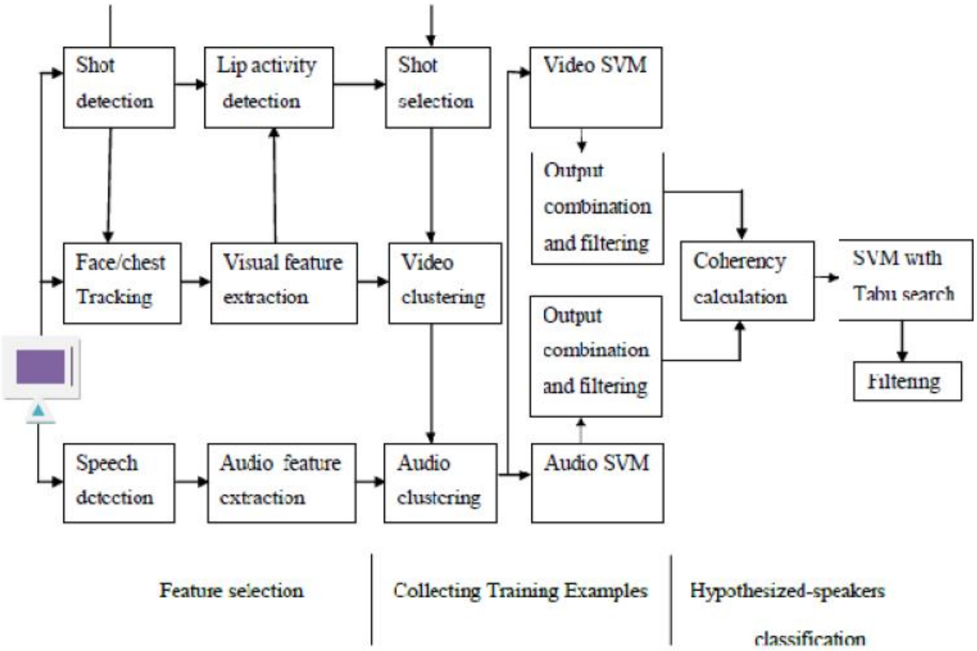

Fig. 1. Architecture of Speaker Diarization in proposed system

talking can be identified by red rectangular boxes and yellow rectangular boxes is discarded for non-talking faces.

Video and audio clustering

Video Clustering: Using the cumulated HSV color histograms for the selected shots, a first (visual) grouping is performed. This one is a hierarchical agglomerative clustering. The distance used to measure the shot similarities is a distance computed on the cumulated color histograms. It is defined as follows:

$$d_x{}^2 = \frac{1}{2}\sum_{i=1}^{b} \frac{(Hx_i - Hy_i)^2}{(Hx_i + Hy_i)}$$

Where H_x and H_y are the number of bins.

Audio Clustering: The audio clustering can be obtained by using SVM classification. The number of clusters can be sampled with the training sets.

SVM Audio Classification of Hypothesized Speakers

The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output. The classification task usually involves training and test sets which consist of data instances. The goal is now to process the remaining parts of the show (that were not selected during the data collection stage and not taken into account in the previous clustering cascade). Note that these remaining parts represent a higher fraction of the content, compared to the shots selected in the previous stage. Practically, we use one-vs-one SVM classifiers, meaning that for the hypothesized speakers *N* obtained earlier, bi-class classifiers are trained. Since the training sets are potentially imbalanced we use a different *C* value for positive and negative training examples, which we will refer to as *C*+ and *C*-respectively.

METHODOLOGY

The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output. The classification task usually involves training and test sets which consist of data instances. Tabu search algorithm is embedded with SVM for clustering and arranging the audio and video samples. Tabu search is a local search method used for mathematical optimization and uses a local or neighborhood search procedure to iteratively move from one potential solution *x* to an improved solution x' in the neighborhood of *x*, until some stopping criterion has been satisfied (generally, an attempt limit or a score threshold). Local search procedures often become stuck in poor-scoring areas or areas where scores plateau. In order to avoid these pitfalls and explore regions of the search space that would be left unexplored by other local search procedures, Tabu search carefully explores the neighborhood of each solution as the search progresses. The solutions admitted to the new neighborhood, N*(*x*), are determined through the use of memory structures. Using these memory structures, the search progresses by iteratively moving from the current solution *x* to an improved solution x' in N*(*x*).

The memory structures used in Tabu search can be divided into three categories:

- Short-term: The list of solutions recently considered. If a potential solution appears on this list, it cannot be revisited until it reaches an expiration point.
- Intermediate-term: A list of rules intended to bias the search towards promising areas of the search space.
- Long-term: Rules that promote diversity in the search process (i.e. regarding resets when the search becomes stuck in a plateau or a suboptimal dead-end).

TS methods generally incorporate with two different strategies to control the efficiency of the search space discovery. These strategies are grouped in two terms: intensification and diversification. The first strategy allows general search of the path to find a best solution. However, if the search is in a region of space for which the solutions are poor or if the general search cannot produce better solutions, the second strategy enables large changes of the solution in order to find quickly another promising region. These two strategies are generally applied alternatively.

Intensification and Diversification strategy

A key element of the adaptive memory framework of TS is to create a balance between search intensification and diversification. Intensification strategies are based on modifying choice rules to encourage move combinations and solution features historically found good. They may also initiate a return to attractive regions to search them more thoroughly. Diversification strategies, however, seek to incorporate new attributes and attribute combinations that were not included within solutions generated previously. In one form, these strategies undertake to drive the search into regions dissimilar to those already examined. It is important to keep in mind that intensification and diversification are not mutually opposing, but are rather mutually reinforcing. Most types of intensification strategies require a means for identifying a set of elite solutions as basis for incorporating good attributes into newly created solutions. Diversification is automatically created in TS (to some extent) by short-term memory functions, but is particularly reinforced by certain forms of longer term memory. TS diversification strategies are often based on modifying choice rules to bring attributes into the solution that is infrequently used. Alternatively, they may introduce such attributes by periodically applying methods that assemble subsets of these attributes into candidate solutions for continuing the search, or by partially or fully restarting the solution process. Diversification strategies are particularly helpful when better solutions can be reached only by crossing barriers or “humps” in the solution space topology.

Performance analysis

By the analysis of previous works, (Bigot *et al.*, 2010) refers the speaker role recognition system correctly 92%, but it will not develop the speaker roles automatically detected in order to achieve audiovisual structuring. The performance of this paper (Richard *et al.*, 2007), the results can be satisfied but more effort can be dedicated to define on unsupervised approach. In the experimental results of (Bozonnet *et al.*, 2010), the complementary results in speaker diarization performance and relative improvements in DER (Diarization Error Rate) of 22% and 17% on the development and evaluation sets. In this paper

(Bendris *et al.*, 2010), lip activity detection was not focused well. In this paper describe about the Tabu search for improve the efficiency and accuracy for searching process. These good results may be explained by very significant improvements on the speaker diarization performance. The expected results can be determined by audio and video output data.

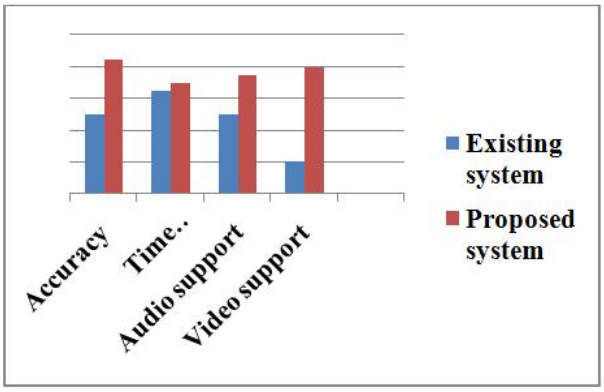

Fig. 2. Comparison of existing and proposed system

Expected results

The impact of the added audiovisual fusion components on the performance of our speaker diarization system can be tested. The speaker diarization method has involved identifying the speakers and non-speakers. In the above graph denotes the comparison of the existing and proposed system. It has accuracy, time complexity, audio support and video support performance criteria.

Conclusion and Futurework

In this article proposed a new multimodal speaker diarization system exploiting speaker recognition methods. The focus has been put on talk-show programs, as such TV content raises challenging research issues, especially from an indexing and event-retrieval perspective for which speaker diarization is a crucial capability. Both audio and video data can be assembled by the association of visual and audio descriptors. The audio and video features can be extracted effectively and processed. Tabu search can be achieved for improved the accuracy for searching method. Two schemes are then considered: an audio-only classification scheme and a parallel audio/visual classification scheme. Time complexity is reduced by using TS-SVM. Finally, the output result can be obtained the optimal solutions. In the future work, the video data can be extracted into number of frames and the frames can be taken much memory storage. So the memory capacity is needed much for process.

REFERENCES

Anguera X., S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 356–370, 2012.

Bendris M., D. Charlet, and G. Chollet, "Lip activity detection for talking faces classification in TV-content," in Proc. Int. Conf.Machine Vision, Hong Kong, China, Dec. 2010.

Bendris, "indexation audio-visuelle des personnes dans un contexte de télévision," Ph.D. dissertation, Telecom ParisTech, Paris, France, 2011.

Bigot B., I. Ferrané, J. Pinquier, and R. André-Obrecht, "Speaker role recognition to help spontaneous conversational speech detection," in *Proc. ACM Workshop Searching for Spontaneous Conversational Speech*, Firenze, Italy, Oct. 2010.

Bozonnet S., F. Vallet, N. Evans, S. Essid, G. Richard, and J. Carrive, "A multimodal approach to initialization for top-down speaker diarization of television shows," in *Proc. European Signal Processing Conf.*, Aalborg, Denmark, Aug. 2010.

Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin, "A Practical Guide to Support Vector Classification, April 15, 2010.

Dielmann, "Unsupervised detection of multimodal clusters in edited recordings," in *Proc. Multimedia Signal Processing*, Saint-Malo, France, Oct. 2010.

Félicien Vallet, Slim Essid, and Jean Carrive, " A Multimodal Approach to Speaker Diarization on TV Talk-Shows", *IEEE Trans. Multimedia*, vol. 15, no.3, pp. 509-520 , APRIL 2013.

Friedland G., H. Hung, and C. Yeo, " Multimodal Speaker diarization of real world meetings using compressed domain video features," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, Taipei, Taiwan, Apr. 2009.

Hubert Cardot, "Tabu Search Model Selection for SVM" International Journal of Neural Systems Special Issue on Issue's Topic.

Richard G., M. Ramona, and S. Essid, "Combined supervised and unsupervised approaches for automatic segmentation of radiophonic audio streams," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, Honolulu, HI, USA, Apr. 2007. http://diarization.icsi.berkeley.edu/diarization/

The NIST Rich Transcription 2009 (RT'09) Evaluation, NIST, 2009. (Online). Available: http://www.itl.nist.gov/iad/mig/tests/rt/ 2009/ docs/rt09-meeting-evalplan-v2.pdf.

Vallet F., S. Essid, J. Carrive, and G. Richard, "Robust visual features for the multimodal identification of unregistered speakers," in *Proc. Int. Conf. Image Processing*, Hong Kong, China, Oct. 2010.

****