Speaker Change Detection for News Audio: A Practical Approach

MsUzma Muzzaffar[1], Dimple Goyal[2], Muheet Ahmed Butt[3]
[1]Student, Dept. of Electronics and Communications, Swami Devi Dyal Institute Of Engineering & Technology, Kurukshetra University, Kurukshetra.
[2]Assistant Professor, Department Of Electronics And Communications, Swami Devi Dyal Institute of Engineering & Technology, Kurukshetra University, Kurukshetra.
[3]Scientist, PG Department of Computer Sciences, University Of Kashmir, Srinagar.

Abstract - Speaker change detection is imperative approach for automatic division process when various speakers communicate at a same case of time. The speech information is decayed into different homogeneous sections with each segment contains the audio information of a single speaker. Existing methodologies for speaker change detection complete disparity of the dispersion of the audio information in a computerized frame prior and then afterward a speaker change point. In this paper, we propose speaker segmentation and clustering based strategy for speaker change identification. Speaker diarisation (or segmentation) is the way toward dividing an information sound stream into homogeneous sections as per the speaker character. Speaker diarisation is a mix of speaker division and speaker grouping. Highlights removed from the sound information around the speaker change focuses are utilized as positive cases. Highlights separated from the information between the speaker changes focuses are utilized as negative illustrations. The positive and negative illustrations are utilized as a part of preparing a help vector machine for speaker change detection.

Introduction

This proposed research tries to investigate a scope of accessible strategies for Speaker Change Detection SCD and analyze them for use in sound altering. Sound altering includes a fairly dreary procedure of acclimation with the individual portions of the media content under observation. The intend is that the previously mentioned and available techniques can ease and unscramble this procedure, subsequently enabling Digital Audio Workstations, DAWs, by including mechanized speaker partitioning an input audio stream into various homogeneous segments.

Ideally a Digital Audio Workstations utilizing speaker segmentation would have the capacity to look inside sound records for nonstandard state data, here referring to themes, speakers, situations, working environment and so on. This proposed research will however concentrate basically on Speacker Change Detection, as it expands on the information assembled from the formation of Castsearch [1], a setting based Spoken Document Retrieval (SDR), a standard web crawler. Amid the production of Castsearch, Jørgensen et al. planned a framework for sound order [2]. This arrangement framework incorporates the classes; Speech, music, clamor and hush. To clear up this proposal will focus exclusively and extensively on sound constituent.

SCD is the procedure of locating the speaker to speaker changes in an audio stream under observation. The proposed research will explain the methods and procedures found in the field of SCD and describe their general application is this proposed research. This proposed research will also touch on speaker clustering. The absolute goal of the proposed research is to hypothesize a set of speaker change-points by comparing samples before and after a potential change-point at regular intervals of time.

Methods for Speaker change detection

Speaker change detection methods used in this project can be further grouped into three subgroups comprising of Gaussian Processes, vector quantization and Direct density-ratio estimation.

The first subgroups are the distance measures between different multivariate Gaussians trained on data before and after the potential change-point. These include the Kullback-Leibler Distance, here termed KL distance or simply KL, and a simplification of it, the so-called Divergence Shape Distance, DSD, which focuses solely on locating covariance changes.

The second subgroup is the Vector Quantization, VQ is an approach which integrates a variety of different approaches to uncover the underlying structure of a dataset under observation through iteratively improved guesses. These improved guesses are in the form of a much smaller amount of typical data, the difference is then measured in the total movement of this representative data under observation and termed Vector Quantization Distortion, VQD.

The magnitude of this number is correlated with the likelihood of a change-point at that moment. These metrics therefore need to be thresholded to yield definite forecasts, rather than a smooth scale of possibilities. These thresholds will be defined relative to a smoothed version of the metric itself

Speech Overlapping

Since real dialogue does not always conform to the simple model of speaker turns the possibility of overlapping speech segments is a obligation and the definition for 'babble noise' is unclear in this sense. Overlapping speech naturally spreads a speaker change over time domain, this may even haze the speaker change to insignificance and a smooth transition to a different speaker altogether is a real liability. The methods discussed above can ease this issue assuming the notion of speaker turns remains valid. This issue naturally lowers the precision of the model, in a sense this issue will be regarded as a single speaker in speech noise only. The data used in this proposed research does not contain overlapping speech in the dataset therefore its imagined consequences are purely theoretical.

Segment clustering of Speaker

The segments can be clustered when the speech has been separated into speaker turn segments and are saved in the matrix format. This provides a conjecture as to how many speakers are present in the speech dataset. The performance of this step may get impacted by the noise in the background environment and other such limitations which are definitely a challenge in the proposed research. Hierarchical Clustering [4], AHC, has been compared and the general concept is to start by assuming that every speaker turn is a unique person. In the proposed research experiment we have presumed every speaker turn is a unique person. In the proposed algorithm for Speaker Change Detection an iteratively combining the most similar segment is done until only there are two segments remaining for observation. The number of speakers is established where the amalgamation of the segment parts are not similar.

Review of Literature

Speaker segmentation has its utility in a most of the speech based applications related to audio and/or video document processing, and information retrieval. The most significant effort in the Rich Transcription domain comes directly from the internationally competitive RT evaluations, sponsored by the National Institute of Standards and Technology (NIST) in the Unites States [1]. In the said research ,the researchers have briefly outlined the standard bottom-up and top-down approaches as well as two recently proposed alternatives: one based on information theory; and a second one based on a non parametric Bayesian approach has been clearly implemented.

In addition to that, some other research works propose sequential single-pass segmentation and clustering [2][4][5], although the researches performance tends to fall short of the state-of-the-art. Various initializations have also been studied and, where as some have investigated k–means clustering, many systems use a uniform initialization, where the audio stream is divided into a number of equal length abutted segments. This simpler approach generally leads to equivalent performance and is covered in the research [6]. Finally, the recently proposed speaker binary keys have been successfully applied to speaker segmentation in meetings of people [7] with similar performance to state-of-the-art systems but also with considerable computation savings (running in around 0.1 times real-time). Speaker binary keys also called as small binary vectors computed from the acoustic data using a universal background model (UBM)-like model.

Most state-of-the-art speaker segmentation engines unify the segmentation and clustering tasks into one step. In these systems, segmentation and clustering are performed hand-in-hand in one loop. Such a method was initially proposed by ICSI for a bottom-up system and has subsequently been adopted by many others [8], [9], [10], [11][12][13]. For top-down algorithms it was initially proposed by LIA [14] as used in their latest system [15].

Methodology

Two approaches have been proposed by Meigner .These are based on the concept of diarisation. He has discussed both its merits and demerits. In the first one, step by step diarisation, the diarisation is done using segmenter and activity detector and speaker clustering. In the other one, to renew clustering algorithm the segmentation and clustering is performed iteratively. This the integrated approach.

Figure 1 below indicates a block diagram of the proposed system which follows the step-by-step segmentation approach

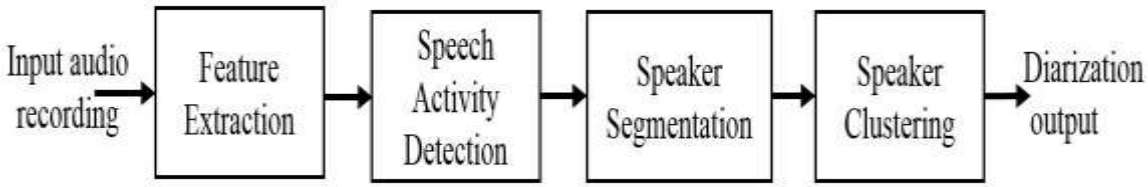


Figure 1: Block diagram of proposed MATLAB system

Speech Activity Detection

Speech activity detection is the process where speech is separated from non-speech in a sound account. Some signal parameters must be registered from the time signal after which they are sent to the discourse recognizer. the spectral shape resolving time for each is 10 ms. The ASR community witnessed some other important developments in the 1980sas well. For SAD in diarisation, false alarm speech badly affects the clustering process and contaminates the speaker models during clustering and segmentation. In speech activity detection, we make an attempt to classify all the sounds which are present in the recording.

Speech Activity Detection Algorithm

A model based classifier is the speech activity detector used. Using an energy based bootstrapping, silence is removed and then iterative classification is done. Next, music and other audible non-speech are identified from the audio recording.

Silence Removal

A confidence value to each frame for silence class and speech classe is given by bootstrap segmentation. Using a Gaussian mixture of size 4 over the 60 dimensional feature spaces is trained using bootstrap silence model. With the same size of high confidence speech frames speech model is also trained

Each frame is classifies into silence and speech class. These high confidence frames are then used to train the speech and silence models for next iteration. As iteration number increases, the number of 60 dimensional Gaussians are increased until a maximum which are used to model the speech and silence GMMs. The results in removal of silences and pauses, having high energy non-speech, also called audible non-speech for example jingles or music are classified as speech. There MFCCs and energy for music resemble speech more than that of silence.

Music removal

In 2005[16] introduced a model fitting based music v/s speech classifier that reported a classification accuracy of 95%. The authors pre-segmented the audio recording into chunks of 1s and extracted 50 feature vectors over 20ms windows. These feature vectors were :

1) short time energy.
2) zero crossing rate

The music speech discriminator[16] fails when speech and music are present together. To prevent loss of informative speech, we used the output of the classifier as a bootstrap segmentation where models for music as well as speech are trained from high confidence frames of both classes. The same iterative classification which is done in silence removal is also done where the speech and music classes are refined to discard music segments only. During the iterative classification step, here we do not use the short time energy. After the short time energy was neglected the speech with music as background classified as music was recovered to the speech class.

Figure 2 provides the speech Activity detection for music, sound and silence in the dataset.

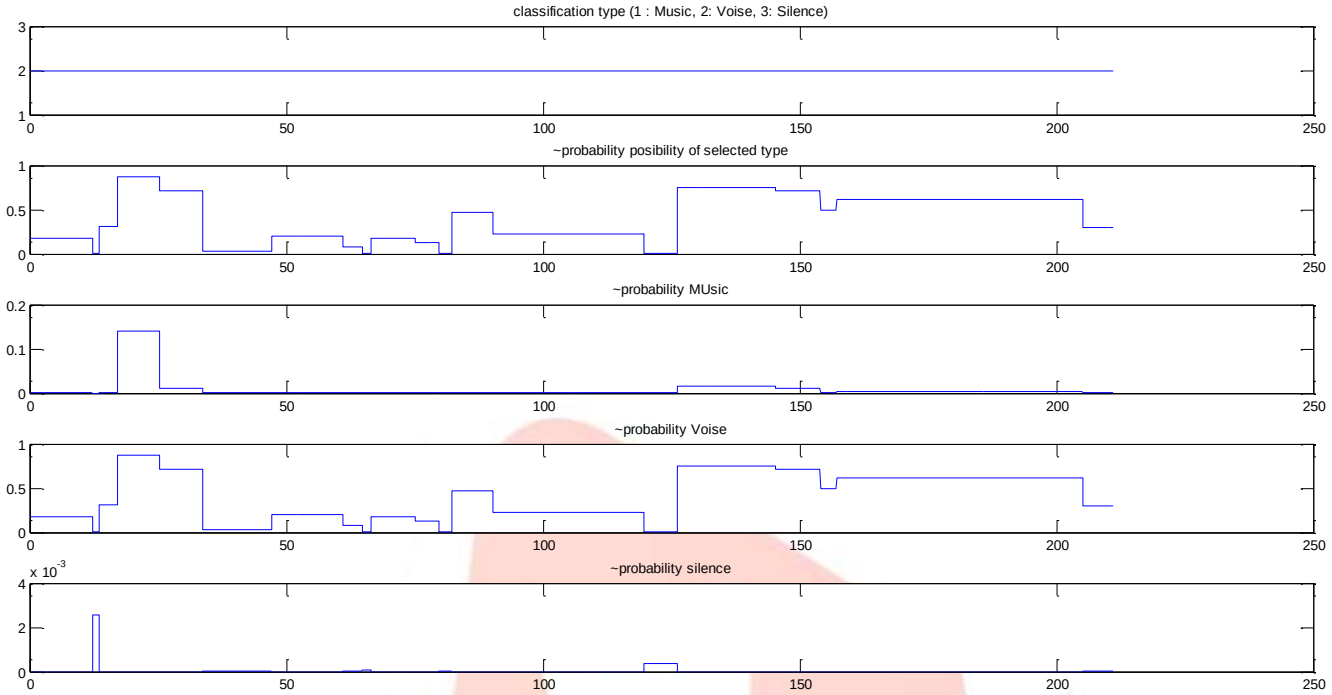


Figure 2: Speech Activity Detection for Music, Voice and Silence

Speech Activity Detection Confidence Measures

The high confidence silence frames are the frames with 20% lowest energies whereas the frames with 10% highest energies are speech with a high confidence. For the training of the GMMs, only these frames are used. For the music removal, the main thing is to rake out from the frames that have speech with music as background but that are classified as non-speech.

Evaluation of Speech Activity Detection

The TIMES NOW dataset was used to tune parameters for silence as well as music removal. These parameters were used to obtain results on the CNBC AWAAZ dataset as well. The two blocks of silence and music removal are decoupled, the set of system parameters is chosen for the former and the output from this block and is then fed to the music removal. Hence the evaluation of the music removal is done separately. In the first step silence is removed from the whole recording using bootstrapping that is energy based and then iterative classification is done . In the second step, music and other audible non-speech are identified from the recording.

Evaluation on the Times Now dataset of duration

The dataset was used as a development set to tune parameters for both silence and music removal.The data source is recording of the news channel TIMES NOW in MP4 format that is of 2min 45 seconds. It is converted into .wav form to use it in MATLAB .These parameters were used to obtain results on the CNBC AWAAZ dataset as well. Most of the time more than one person is speaking. The music evaluation is done separately. The segmentation and classification processes are shown in figure 3.

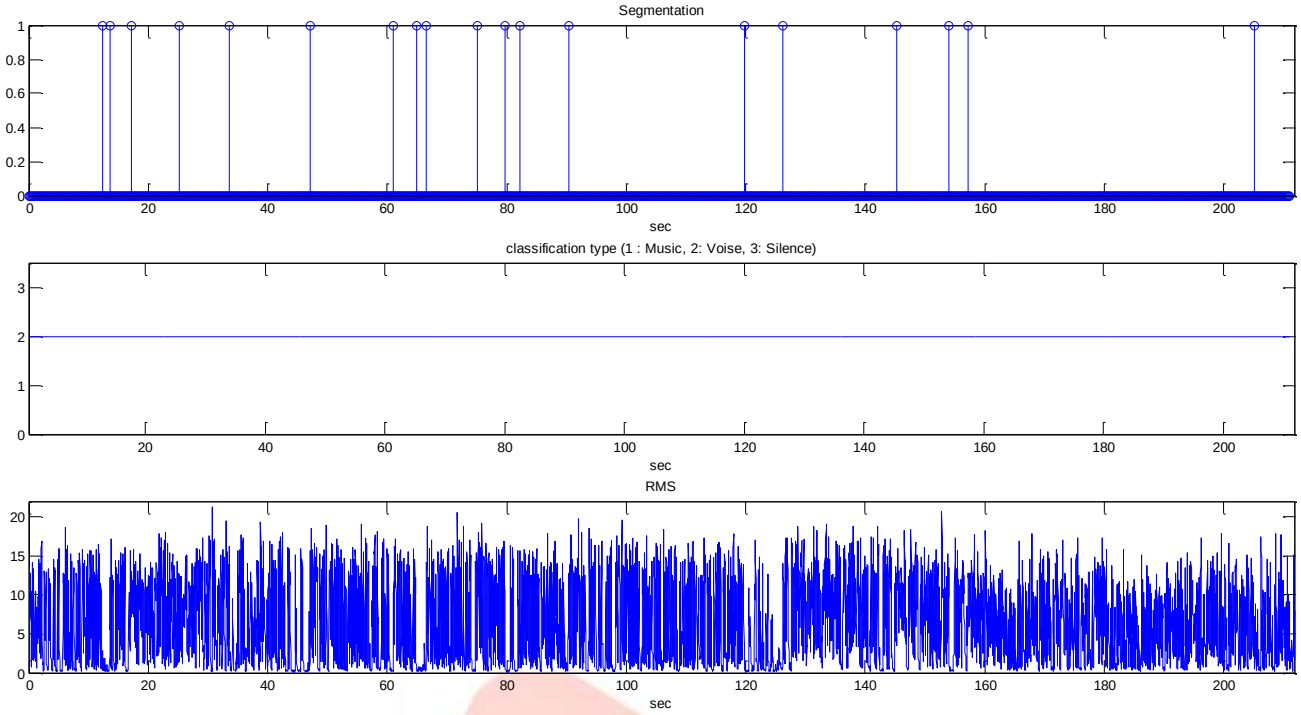
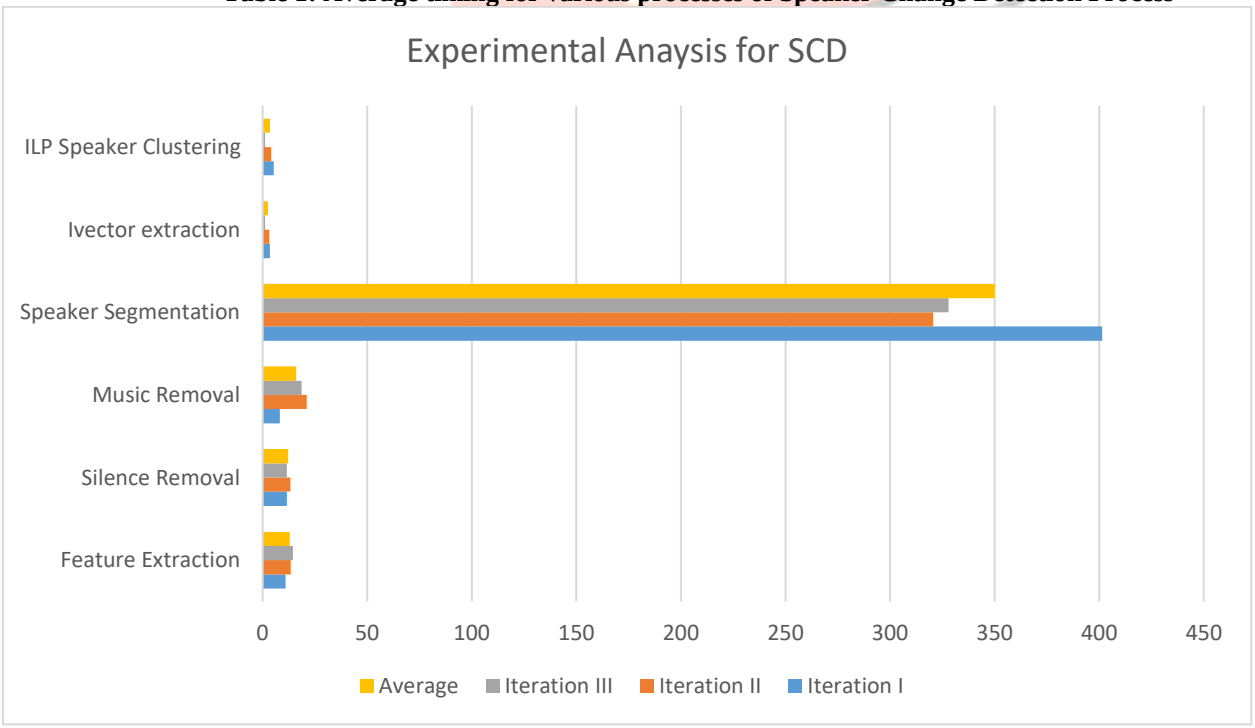


Figure 3: Segmentation, Classification and RMS for the Speech

The experimental analysis for various steps is shown in the table below

	Iteration I	Iteration II	Iteration III	Average
Feature Extraction	**10.9503**	**13.4434**	**14.4414**	**12.94503**
Silence Removal	**11.597**	**13.2643**	**11.5726**	**12.14463**
Music Removal	**8.1441**	**21.1458**	**18.5835**	**15.9578**
Speaker Segmentation	**401.3841**	**320.5587**	**327.9511**	**349.9646**
Ivector extraction	**3.4128**	**3.2131**	**1.1077**	**2.577867**
ILP Speaker Clustering	**5.3817**	**4.053**	**1.1691**	**3.5346**

Table 1: Average timing for various processes of Speaker Change Detection Process



Graph 1: Experimental Analysis for Speaker Change Detection

Choice of speaker model

In this system the speaker models that have been used are Gaussian Mixture models and i-vector models. These have been widely used in speech verification systems under observation. The GMM is a probabilistic model on the feature space and the the i-vector

systems have become the state-of-the-art technique in the speech verification systems . They provide a unique way of reducing the large-dimensional input data to a small-dimensional feature vector while retaining most of the relevant information

i-vector extraction

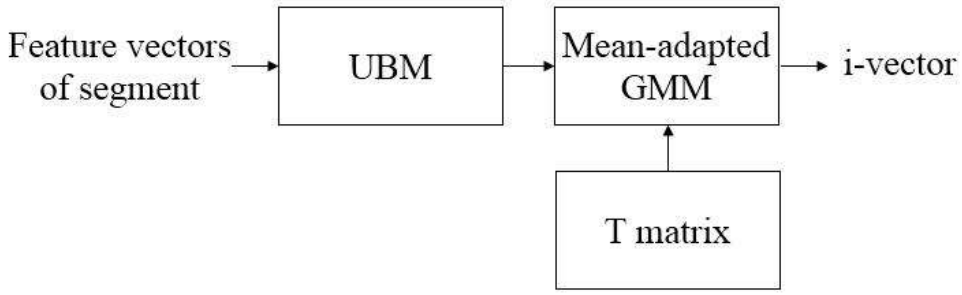
In order to obtain relevant speaker features for the proposed research, the vectors are analyzed towards characteristic factors. Thereby, a factor analysis model is iteratively trained and the factor analyzed features are referred to as identity-vectors (i-vectors) are obtained. So as to efficiently obtain the i-vectors, first a speech Universal Background Model (UBM) is trained on a training data set. The UBM is a GMM with large number of Gaussians so that it captures all possible inconsistencies in speech under observation in the feature space.

The Total Variability space is a subspace of the GMM super space. It seizes all the channel and speaker related information. T is called the low rank matrix . For this system, the matrix T is trained using the speaker labelled dataset used for UBM training. The i-vector of the segment is the projection of the GMM supervectors onto the Total Variability subspace.

$$m = M + T x$$

where M is the UBM supervector, m is the mean-adapted GMM supervector of the segment.

Figure 4: Extraction of i-vectors



Two distance metrics have been tested for measuring similarity between i-vectors -the cosine similarity metric (1.1) and the Mahalanobis distance metric (1.2) where W is the within class covariance matrix determined from the n training i-vectors from S speakers detailed in 1.3.The Mahalanobis distance is hence also called within class covariance normalization (WCCN). In equation 1.3 of WCCN computation for the Mahalanobis distance, the vectors $\overline{w^s}$ are mean of the n_s i-vectors of speaker s

$$D(x,y) = 1 - \frac{x^T.y}{||x||.||y||} \quad (1.1)$$

$$D(x,y) = (x-y)^T W^{-1}(x-y) \quad (1.2)$$

$$W = \frac{1}{n}\sum_{s=1}^{S}\sum_{i=1}^{n_s}(w_i^s - \overline{w^s})(w_i^s - \overline{w^s})^T \quad (1.3)$$

CONCLUSION

This paper we have investigated, implemented, contrasted and combined a wide range of methods for speaker change detection using segmentation, and subsequently selected a method on which novel improvements have been implemented. The methods used are mostly generic algorithms for various processes of speaker change detection. The investigated speaker change detection methods were drawn from the fields of vector quantization, Gaussian processes and relative density-ratio estimation. These methods mainly included an optimized K means algorithm, the Kullback-Leibler distance, KL, and Relative unconstrained Least-Squares Importance Fitting, RuLSIF, respectively.

Experiments were performed on TIMESNOW datasets were prepared and we have got good results for simple dataset. The data source is recording of the news channel TIMES NOW of MP4 format that is of 2 min 45 seconds which is converted into wav format for research input for the MATLAB code. Speech activity detection has been done using two stages of Speech activity Detection i.e silence detection and removal and then music detection and removal. Short term energy and zero crossing rate have been used in the proposed research as features to evaluate the basic characteristics of speech .i-vector speaker change models provide a low dimensional representation of the speaker information in comparison to GMM speaker models and also offer a computational advantage.

References

[1] "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-eval-plan-v2.pdf

[2] M. Kotti, E. Benetos, and C. Kotropoulos, "Computationally efficient and robust BIC-based speaker segmentation," IEEE Trans. Audio, Speech, Lang. Process., vol. 16, no. 5, pp. 920–933, Jul. 2008.

[3] Song Liu, Makoto Yamada, Nigel Collier and Mashashi Sugiyama, Change point Detection in time-series data by relative density ratio estimation, Neural Networks,2013.

[4] X. Zhu, C. Barras, L. Lamel, and J.-L. Gauvain, "Multi-stage speaker diarization for conference and lecture meetings," in Proc. Multimodal Technol. Perception of Humans: Int.Eval.WorkshopsCLEAR2007and RT 2007, Baltimore, MD, May 8–11, 2007, Revised Selected Papers, Berlin, Heidelberg: Springer-Verlag, 2008, pp. 533–542.

[5] S. Jothilakshmi, V. Ramalingam, and S. Palanivel, "Speaker diarization using autoassociative neural networks," Eng. Applicat. Artif. Intell., vol. 22, no. 4-5, pp. 667–675, 2009.
[6] X.Anguera, C.Wooters, and J.Hernando, "Robust speaker diarization for meetings: ICSI RT06s evaluation system," in Proc. ICSLP, Pittsburgh, PA, Sep. 2006.
[7] X. Anguera and J.-F. Bonastre, "Fast speaker diarization based on binary keys," in Proc. ICASSP, 2011.
[8] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," in Multimodal Technologiesfor Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007,Baltimore, MD, USA, May 8–11, 2007, Revised Selected Papers, Berlin, Heidelberg: Springer-Verlag, 2008, pp. 509–519.
[9] X. Anguera, C. Wooters, B. Peskin, and M. Aguilo, "Robust speaker segmentation for meetings: The ICSI-SRI spring 2005 diarization system," in Proc. NIST MLMI Meeting Recognition Workshop, Edinburgh, U.K., 2005.
[10] D. A. V. Leeuwen and M. Kone˘cný, "Progress in the AMIDA speaker diarization system for meeting data," in Proc. Multimodal Technol. for Percept. of Humans: Int. Eval. Workshops CLEAR 2007 and RT 2007, Baltimore,MD,May8–11,2007,RevisedSelectedPapers,Berlin,Heidelberg: Springer-Verlag, 2008, pp. 475–483.
[11] G.Friedl and, O. Vinyals, Y.Huang, and C.Muller,"Pro sodic and other long-term features for speaker diarization,"IEEETrans .Audio, Speech, Lang. Process., vol. 17, no. 5, pp. 985–993, Jul. 2009.
[12] J. Luque, X. Anguera, A. Temko, and J. Hernando, "Speaker diarization for conference room: The UPC RT07s evaluation system," in Proc. Multimodal Technol. Perception of Humans: Int. Eval. Workshops CLEAR 2007 and RT 2007, Baltimore, MD, May 8–11, 2007, Revised Selected Papers, Berlin, Heidelberg: Springer-Verlag, 2008, pp. 543–553.
[13] J. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multiple distant microphone meetings: Mixing acoustic features and interchannel time differences," in Proc. Interspeech, 2006.
[14] S. Meignier, J.-F. Bonastre, and S. Igounet, "E-HMM approach for learning and adapting sound models for speaker indexing," in Proc. Odyssey Speaker and Lang. Recognition Workshop, Chania, Creete, Jun. 2001, pp. 175–180.
[15] C.Fredouille,S.Bozonnet,andN.W.D.Evans,"TheLIA-EURECOM RT'09 speaker diarization system," in Proc. RT'09, NIST Rich Transcription Workshop, Melbourne, FL, 2009.
[16] Costas Panagiotakis and George Tziritas. A speech/music discriminator based on rms and zero-crossings. Multimedia, IEEE Transactions on, 7(1):155{166, 2005