Comparison of Diarization Tools for Building Speaker Database

Eva KIKTOVA, Jozef JUHAR

Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University of Kosice, Park Komenskeho 13, 041 20 Kosice, Slovakia

eva.kiktova@tuk.sk, jozef.juhar@tuke.sk

DOI: 10.15598/aeee.v13i4.1468

Abstract. *This paper compares open source diarization toolkits (LIUM, DiarTK, ALIZE-Lia_Ral), which were designed for extraction of speaker identity from audio records without any prior information about the analysed data. The comparative study of used diarization tools was performed for three different types of analysed data (broadcast news - BN and TV shows). Corresponding values of achieved DER measure are presented here. The automatic speaker diarization system developed by LIUM was able to identified speech segments belonging to speakers at very good level. Its segmentation outputs can be used to build a speaker database.*

Keywords

ALIZE-Lia_Ral, DiarTk, LIUM_SpkDiarization, speaker diarization.

1. Introduction

The current state of the science and knowledges has a strong influence on our society. One of the highly attractive research topic is speech-to-text transcription, where the fusion of different theoretical knowledges, experiments and finally prototype realizations lead to the complex system.

Over recent years, a speaker diarization has become an important key technology for many tasks such as navigation, retrieval or segmentation to the homogeneous regions of audio data [1].

An audio diarization is the process of annotating an input audio stream with information that attributes to segments of signal energy to their specific sources such as speakers, music, background noise sources, and other signal source characteristics. The diarization can be also used in the speech recognition, facilitating the searching and indexing of audio archives, increasing the richness of automatic transcriptions and for the enhancement their readability [2]. The effective diarization tool can also be used for the automatic database creation from large amount of acoustic data.

The speaker diarization, the "who spoke when" task, consists in annotating recordings with labels that represent speakers. This task is performed without any prior information: neither the number of speakers, nor their identities, nor samples of their voices are available [3].

There are two main kinds of clustering strategies, which can be used in a diarization system. The first is called bottom-up, also known as an agglomerative hierarchical clustering (AHC). The algorithm starts in splitting the full audio content in a succession of clusters and progressively tries to merge the redundant clusters in order to reach a situation where each cluster corresponds to a real speaker. The Bayesian Information Criterion (BIC), Kullback-Leibler (KL) or T_s based metric can be applied as a stop criteria [1].

The second clustering strategy is called top-down and starts with one single cluster for all the audio content and tries to split it iteratively one-by-one until reaching the number of clusters equal to the number of speakers. Previous mentioned stopping criteria can be applied to terminate the process or it can continue until no unlabelled data remain. The bottom-up approach is more popular.

This paper compares open-source diarization toolkits: LIUM_SpkDiarization, DiarTk and ALIZE-Lia_Ral. The speaker segmenation and clustering are based only on the audio information (i.e. without any additional information such as a number of speakers, etc.). Obtained speaker tags don't represent identities but abstract labels. Three different data types for which the mentioned toolkits have been used are re-

ported in this paper. One are short broadcast news (BN), then speech of main anchormen of daily BN and the last type is SUS TV-show.

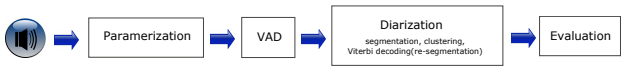

Fig. 1: Principal block scheme of diarization system.

2. Diarization System

Figure 1 shows the main modules of an overall diarization system. It composed of parametrization, possible VAD/SAD detector, diarization and evaluation module. The audio processing starts with the parametrization module, that is responsible for the features extraction. The next module performs VAD (Voice Activity Detection) in better case SAD (Speech Activity Detection). Ideally, only speech segments are processed by the diarization module. These speech data are divided to clusters and finally after the whole diarization process (according to the diarization tool) only clusters corresponding to speakers are provided. The diarization output is then evaluated in the last module.

2.1. Parametrization

Popular features used in such systems are Mel-Frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction coefficients (PLP), Linear Frequency Cepstra Coefficients (LFCC), fundamental frequency, energy, etc., [15], with their temporal extensions such as delta and acceleration coefficients. Depending on the used diarization system it is possible to omit some features (c_0 in the LIUM) or used a combination of different features (e.g. DiarTK [7]).

2.2. Speech Activity Detection

The Speech Activity Detection (SAD) plays a very important role in the whole diarization process for two reasons [1], [2]. The first is the impact on the speaker diarization performance metric, namely the Diarization Error Rate (DER). The evidence of non-speech sounds increased a diarization error. The second follows from the fact that non-speech segments can disturb the acoustic modelling of speaker dependent models and make them less discriminant. An initial approaches for diarization try to solve SAD on the fly, i.e. non-speech clusters were a by-product of the diarization.

2.3. Diarization

The currently used diarization tools are LIUM_SpkDiarization [5], DiarTk [7], AudioSeg [8], SHoUT [9], ALIZE [10] and tool developed by IRIT SAMoVA group [11]. The main diarization processes are segmentation, clustering and realignment. Depending on the particular tool some differences can be found.

An audio stream is segmented to the speech and non-speech frames [7]. Speech segments are then processed by the Hierarchical Agglomerative Clustering (HAC) in which segments belonging to the same speaker are clustered together. Viterbi decoding (re-segmentation) is performed to generate a new segmentation realigned on the speaker boundaries. LIUM_SpkDiarization system finally performed another HAC using Cross-Likelihood Ratio (CLR) (classical or normalized) or Integer Linear Programming (ILP) proposed in [12], where i-vectors were used to model and measure the similarity between clusters. The diarization output contains the time stamps of segments that belong to the each recognized speaker.

2.4. Evaluation

For computing a Diarization Error Rate (DER) on the speech segments [6], three error types have to be defined:

- The confusion error - the system-provided speaker tag and the reference do not match through the mapping.
- The miss error - speech is present in the reference but no speaker is present in the system hypothesis.
- The false alarm error - speech is incorrectly detected by the system.

The used speaker diarization systems were evaluated by the NIST evaluation procedure for computing DER using rttm files (perl script: md-eval-v21.pl):

$$\mathrm{DER} = \frac{\text{confusion} + \text{miss} + \text{false alarm}}{\text{total reference speech time}}. \qquad (1)$$

3. Used Diarization Tools

The brief theoretical description of analysed diarization tools can be found below, namely LIUM_SpkDiarization, DiarTk and ALIZE-Lia_Ral.

3.1. LIUM_SpkDiarization

The open-source toolkit *LIUM_SpkDiarization* [3], [5] was developed from a previous speaker segmentation tool, *mClust* [16] in C++ by LIUM for the French ESTER evaluation campaigns in 2005 and 2008. This toolkit was designed to processing TV shows and radio broadcast. It analyses the input audio stream, performs diarization and identifies homogeneous segments belonging to the same speaker without any prior information about an audio content e.g. number of speakers.

The diarization system provided by LIUM starts its processing by the computation of 13 MFCC with c_0 using Sphinx tools (`http://cmusphinx.sourceforge.net/`). This configuration of features is not used through whole diarization but for example in the Viterbi decoding phase, c_0 is removed and first order derivative are added to the feature vectors.

Two phases speaker segmentation is based on GLR (Generalized Likelihood Ratio) for the identification of instantaneous change points and BIC (Bayesian Information Criterion) distance metric for the fusion of consecutive segments belonging to same speaker.

BIC hierarchical clustering merges two closest clusters until BIC distance is positive. In the segmentation and clustering phase speakers are modelled by Gaussian distribution with full covariance matrix.

Viterbi decoding is performed to adjust segment boundaries using GMMs as speaker models. Feature vectors are modified as was described above. The speech/non-speech segmentation and music & jingle regions removal is done in this phase [5]. The decoding uses 8 GMMs corresponding to 2 silences (wide and narrow band), 3 types of wide band speech (clean, over noise or over music), 1 narrow band speech, 1 music and 1 jingle. The GMMs contain 64 diagonal Gaussians trained by EM-ML on ESTER data [4].

LIUM_SpkDiarization system finally performs another HAC using normalized Cross-Likelihood Ratio (CLR) or Integer Linear Programming (ILP) proposed in [12], where i-vectors were used to model and measure the similarity between clusters [5].

Diarization outputs were converted to the rttm file format for obtaining Diarization Error Rate DER (%) according to the NIST evaluation procedure.

3.2. DiarTk

The open source toolkit *DiarTk* [7] was developed for multi-stream speaker diarization tasks under GPL licence. It was developed in C++.

The diarization process consists of three main operations. The first is a segmentation into homogeneous regions, then an agglomerative clustering is performed where segments are grouped according to the speaker. Finally a Viterbi realignment represents a diarization output, in which speaker segment boundaries are refined. DiarTk is able to handle with multi feature streams (MFCC, Time Delay of Arrivals - TDOA, Modulation Spectrum - MS, Frequency Domain Linear Prediction features - FDLP) [13], [14]. The main difference against a conventional diarization system is in the speaker modelling technique: DiarTk employs the non-parametric clustering and realignment based on the agglomerative Information Bottleneck principle (it does not use GMM speaker modelling) [7].

The diarization algorithm can be briefly described for one feature stream as follows [7]: It starts with the feature extraction, then the audio segmentation tool ($IBfeat$) performs speech/non-speech segmentation, non-speech frames elimination, background GMM estimation and computation of relevance variable distributions $p(Y|X)$ as a weighted sum of individual distributions $W_i p(Y|X_i)$, where X represents speech segments and Y relevance variables information.

The second $aIBclust$ tool performs hierarchical agglomerative clustering, where speech segments X associated with the relevance variable distributions $p(Y|X)$ are clustered in to clusters C.

The last module $IBrealign$ is responsible for the Viterbi realignment of speaker boundaries. This process as well depends only on the relevance variable distributions $p(Y|X)$.

The diarization output is in the rttm format compatible to be scored by the NIST evaluation module.

3.3. ALIZE-LIA_RAL

This complete diarization toolkit in C++ is an open-source distributed under GPL license. The package ALIZE [5], [10] provides the basic operations required for handling configuration files and features, matrix operations, error handling, etc., and LIA_RAL package performs several tasks including language recognition, speaker recognition, diarization/segmentation, VAD, etc.

In the acoustic segmentation step, the VAD classifies the audio content in the following predefined classes: speech, music, music+speech, or telephone. It is realized by build-in GMMs, or can be used speech non-speech GMM detection based only on the energy. VAD models can be created or downloaded from official webpage. Then the segmentation and speaker clustering are performed by using evolutive HMMs (e-HMM).

The additional segmentation is done in order to refine the previous speaker segmentation output and to remove irrelevant speakers (i.e. speakers with a low number of assigned frames). In this case the algorithm creates a new HMM models generated from the previous segmentation output and then applying an iterative speaker model training/Viterbi decoding loop [10].

An optional purification step can be integrated within LIA_RAL toolkit. The purification step is then realized between the segmentation and resegmentation phase. It eliminates the spurious segments. The diarization output is in the rttm file format.

4. Database

In our experimental work all used data had wav format, mono, 16 kHz sampling frequency. Three different types of acoustic data were diarized (see the Fig. 2):

- Short broadcast news included jingles, speech (anchorman, redactors and respondents), phone speech and background music. The average short news duration was about 5 minutes.
- Second diarized data included two anchormen speech during main daily broadcast news. They contained only main anchorman (primary sound) a cross talk of co-anchorman. The average news duration was about 1 hour.
- The last type of diarized data were TV shows SUS ("Court hall"). They contained advertisement, jingles, several speakers, overlapped speech and background music. The average show duration was about 1 hour. This type of sound data is characterized by dynamic dialogues between participants of the court hearing.

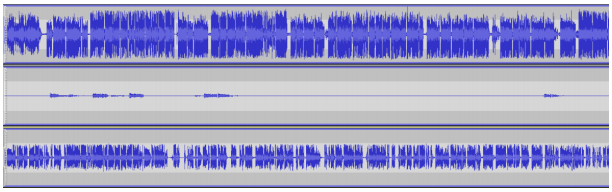

Fig. 2: Examples of analyzed sound data (first track - short BN, second track - anchormen daily BN, third track - TV show - SUS).

5. Results

Achieved results for three different data types can be found in the Tab. 1. As was mentioned, for DER (%) computation the rttm file format is required. Obtained results are grouped according to the analysed data. Average DER (%) was computed for each data type separately. SUS data in the comparison of both BN data were much more demanding.

Tab. 1: Diarization results.

Data	DER (%)		
	LIUM	Lia_Ral	DiarTk
short_BN_1	13.66	64.30	70.01
short_BN_2	32.53	40.74	46.29
short_BN_3	16.31	46.32	50.61
short_BN_4	19.07	40.88	28.28
short_BN_5	8.24	58.60	60.78
short_BN_6	10.24	39.08	74.36
short_BN_7	32.27	47.07	89.97
short_BN_8	30.00	52.80	84.46
short_BN_9	28.00	23.34	86.76
short_BN_10	11.88	57.18	28.76
Average short_BN	20.22	47.03	62.03
daily_BN_1	3.76	68.44	78.66
daily_BN_2	1.26	50.79	26.51
daily_BN_3	2.26	64.21	8.24
daily_BN_4	3.27	71.38	2.99
daily_BN_5	1.81	82.00	4.13
daily_BN_6	2.66	55.97	29.40
daily_BN_7	4.41	93.52	2.58
daily_BN_8	2.92	51.89	22.55
daily_BN_9	2.26	56.17	18.59
daily_BN_10	7.33	53.70	22.15
Average daily_BN	3.19	64.81	20.98
SUS_01	24.81	64.47	45.76
SUS_02	19.80	20.03	16.36
SUS_03	18.41	12.11	30.94
SUS_04	86.63	46.58	32.63
SUS_05	14.03	30.96	74.42
SUS_06	15.98	35.82	21.34
SUS_07	16.00	18.28	19.86
SUS_08	33.96	39.77	26.02
SUS_09	16.93	39.17	17.05
SUS_10	46.75	28.62	28.51
Average SUS	29.33	33.58	31.29

The majority of errors for short_BNs diarization was caused during external contributions and during telephone or degraded speech for all analysed tools. Generally, this data type achieved high DER values.

Daily_BNs were diarized very good by the LIUM, but in the case of Lia_Ral were achieved high DER values (on average 64.81 %). It was caused by the applied VAD algorithm, that classified the co-anchorman low level speech as a non-speech.

As was mentioned before, SUS data were very challenging, but LUIM_SpkDiarization was able to identified the primary speakers (judge, advocates and another main participants). Other tools (Lia_Ral and DiarTk) achieved worse, but comparable results.

Promising results were achieved mainly for daily_BNs (on average 3.19 % DER) with LIUM, where the correct identification of speakers was at very high level. Some problems were identified in the case of simultaneous speech and during laughter. Such kind of data can be very effective processed by LIUM.

6. Conclusion

In this paper the diarization of three different types of acoustic data were performed. The comparison of obtained DER values was presented. LIUM_SpkDiarization seems to be the effective tool for speaker segmentation and identification for all tested data types.

The speaker database can be created according to the LIUM by using only speakers that have high occurrence in the diarization output. This way an elimination of hazardous segments will be achieved.

An automatic transcription system (ATS) can be built from an effective diarization and an automatic speech recognition system. Then ATS should allow identification of audio document structure and automatic speech recognition with an extraction of the speaker identity.

In the future we would like to perform the diarization also with other type of data (for example meeting speech and lecture recordings).

Acknowledgment

This work is the result of the Project implementation: University Science Park TECHNICOM for Innovation Applications Supported by Knowledge Technology, ITMS: 26220220182 (50 %), supported by the Research & Development Operational Programme funded by the ERDF and by the research project VEGA 1/0075/15 (50 %).

References

[1] ANGUERA, M. X., S. BOZONNET, N. EVANS, C. FREDOUILLE, G. FRIEDLAND and O. VINYALS. Speaker Diarization: A Review of Recent Research. *IEEE Transactions on Audio, Speech, and Language Processing*. 2012, vol. 20, no. 2, pp. 356–370. ISSN 1558-7916. DOI: 10.1109/TASL.2011.2125954.

[2] REYNOLDS, D. A. and P. T. CARRASQUILLO. Approaches and applications of audio diarization. In: *International Conference on Acoustics, Speech, and Signal Processing*. Pennsylvania: IEEE, 2005, pp. v/953–v/956. ISBN 0-7803-8874-7. DOI: 10.1109/ICASSP.2005.1416463.

[3] MEIGNIER, S. and T. MERLIN. LIUM_SpkDiarization: An Open Source Toolkit For Diarization. In: *CMU SPUD Workshop*. Dallas: AMC, 2010, pp. 1–6.

[4] JOUSSE, V., S. PETITRENAUD, S. MEIGNIER, Y. ESTEVE and C. JACQUIN. Automatic named identification of speakers using diarization and ASR systems. In: *International Conference on Acoustics, Speech and Signal Processing*. Taipei: IEEE, 2009, pp. 4557–4560. ISBN 978-1-4244-2353-8. DOI: 10.1109/ICASSP.2009.4960644.

[5] ROUVIER, M., G. DUPUY, P. GAY, E. KHOURY, T. MERLIN and S. MEIGNIER. An Open-source State-of-the-art Toolbox for Broadcast News Diarization. In: *13th Annual Conference of the International Speech Communication Association*. Lyon: ANR, 2013, pp. 1–5.

[6] O. GALIBERT. Methodologies for the evaluation of Speaker Diarization and Automatic Speech Recognition in the presence of overlapping speech. In: *13th Annual Conference of the International Speech Communication Association*. Lyon: ANR, 2013, pp. 1–4.

[7] VIJAYASENAN, D. and F. VALENTE. DiarTk: An Open Source Toolkit for Research in Multistream Speaker Diarization and its Application to Meetings Recordings. In: *12th Annual Conference of the International Speech Communication Association*. Portland: ANR, 2012, pp. 1–5.

[8] GRAVIER, G., M. BETSER and M. BEN. Audioseg: Audio Segmentation Toolkit. *InriaForge* [online]. 2010. Available at: `https://gforge.inria.fr/frs/download.php/file/25187/audioseg-1.2.pdf`.

[9] HUIJBREGTS, M. *Segmentation, diarization and speech transcription: Surprise data unraveled*. Twente, 2008. Thesis. University of Twente. Supervisor: F. M. G. de Jong.

[10] BONASTRE, J.-F., N. SCHEFFER, D. MATROUF, C. FREDOUILLE, A. LARCHER, A. PRETI, G. POUCHOULIN, N. EVANS, B. FAUVE, and J. MASON. ALIZE/SpkDet: a state-of-the-art open source software for speaker recognition. In: *Odyssey: the Speaker and Language Recognition Workshop*. Stellenbosch: ISCA, 2008, pp. 1–8.

[11] EL KHOURY, E., C. SENAC and R. ANDRE-OBRECHT. Speaker Diarization: Towards a More Robust and Portable System. In: *International Conference on Acoustics, Speech and Signal Processing*. Honolulu: IEEE, 2007, pp. 489–492. ISBN 1-4244-0727-3. DOI: 10.1109/ICASSP.2007.366956.

[12] ROUVIER, M. and S. MEIGNIER. A global optimization framework for speaker diarization. In: *Odyssey: the Speaker and Language Recognition Workshop*. Singapore: ISCA, 2012, pp. 1–5.

[13] VIJAYASENAN, D., F. VALENTE and H. BOURLARD. Multistream speaker diarization beyond two acoustic feature streams. In: *International Conference on Acoustics, Speech and Signal Processing*. Dallas: IEEE, 2010, pp. 4950–4953. ISBN 978-1-4244-4295-9. DOI: 10.1109/ICASSP.2010.5495086.

[14] GARAU, G. and H. BOURLARD. Using audio and visual cues for speaker diarisation initialisation. In: *International Conference on Acoustics, Speech and Signal Processing*. Dallas: IEEE, 2010, pp. 4942–4945. ISBN 978-1-4244-4295-9. DOI: 10.1109/TASL.2009.2015089.

[15] FRIEDL, G., O. VINYALS, Y. HUANG and C. MULLER. Prosodic and other Long-Term Features for Speaker Diarization. *IEEE Transactions on Audio, Speech, and Language Processing*. 2009, vol. 17, no. 5, pp. 985–993. ISSN 1558-7916. DOI: 10.1109/TASL.2009.2015089.

[16] FRALEY, C., A. E. RAFTERY, T. B. MURPHY, L. SCRUCCA. mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. *my.ilstu.edu* [online]. 2012. Available at: `http://my.ilstu.edu/~mxu2/mat456/mcluster.pdf`.

About Authors

Eva KIKTOVA was born in Liptovsky Mikulas, Slovakia in 1984. In 2009 she graduated M.Sc. (Ing.) at the Department of Electronics and Multimedia Communications of the Faculty of Electrical Engineering and Informatics at the Technical University of Kosice. In 2013 she recieved Ph.D. at the same department in the field of Telecommunications, where she works as a researcher. Her research is oriented on the field of the acoustic event detection and classification, speaker recognition and speaker diarization.

Jozef JUHAR was born in Poproc, Slovakia in 1956. He graduated from the Technical University of Kosice in 1980. He received Ph.D. degree in Radioelectronics from Technical University of Kosice in 1991, where he works as a Full Professor at the Department of Electronics and Multimedia Communications. He works as a head of the same department. He is author and co-author of more than 200 scientific papers. His research interests include digital speech and audio processing, speech/speaker identification, speech synthesis, development in spoken dialogue and speech recognition systems in telecommunication networks.