SPEAKER DIARIZATION USING DATA-DRIVEN AUDIO SEQUENCING

Houssemeddine Khemiri[1,2], *Dijana Petrovska-Delacrétaz*[2], *Gérard Chollet*[1,3]

[1] Institut Mines-Télécom; Télécom ParisTech; CNRS LTCI
[2] Institut Mines-Télécom; Télécom SudParis ; CNRS SAMOVAR
[3] Department of Electrical & Computer Engineering, Boise State University, Idaho, USA
khemiri,chollet@telecom-paristech.fr,dijana.petrovska@telecom-sudparis.eu

ABSTRACT

In this paper, a speaker diarization system based on data-driven segmentation is proposed. In addition to the usual segmentation and clustering steps, a new module which detects repeated segments between the same shows broadcasted on different dates is added. This process is achieved by using the ALISP-based audio identification system which segments audio data into pseudo-phonetic units. The ALISP segmentation is then used to identify the similar audio segments in TV and radio shows. The system was evaluated during the ETAPE 2011 evaluation campaign and obtained a Diarization Error Rate - DER of 16.23% which was the best result among seven participants.

Index Terms— speaker diarization, ALISP units, data-driven audio sequencing.

1. INTRODUCTION

In a speaker diarization task the goal is to segment an input audio stream into homogenous regions according to speaker's identities in order to answer the question "Who spoke when?". Speaker diarization is a very useful preprocessing step for many audio technologies such as automatic speech and speaker recognition, audio indexing or rich transcription.

We are interested in speaker diarization for TV and radio shows which include various acoustic sources such as studio/telephone speech, music, or speech over music. Speaker diarization relies on a speaker segmentation step, followed by a clustering process. The speaker segmentation step aims to locate the boundaries of speech segments by finding the speaker change or more generally acoustic change points. Speaker clustering is applied to the speech segments that seem to be pronounced by the same speaker.

Several diarization systems have been reviewed by Anguera et al. in [1]. Most of these systems are divided into two categories: bottom-up and top-down approaches. Bottom-up approaches are by far the most common in the literature. Such approaches consist of over-segmenting the audio stream into equal-length segments, training a number of models, and successively merging and reducing the number of clusters until only one remains for each speaker. On the other hand, top-down approaches start with a single speaker model trained on all speech segments, and add new speakers until the stop criterion is reached.

In this paper, a new module based on audio fingerprinting concept is added. We exploit the fact that TV and radio shows keep generally the same structure with same presenters and jingles. We propose to compare the show to be segmented with the same show broadcasted before in order to find the common audio segments. This operation is performed via audio fingerprinting which involves the extraction of a fingerprint for each audio document stored in a reference database. An unlabeled audio excerpt is identified by comparing its fingerprint with those of the reference database.

An audio fingerprint is a compact content-based signature that summarizes an audio recording. In the proposed system, the audio fingerprint is extracted from audio data using a data-driven audio sequencing (segmentation) based on ALISP (Automatic Language Independent Speech Processing) tools. These tools were first developed for very low bit-rate speech coding [2], and then successfully adapted for other tasks such as speaker [3], and language recognition [4], and audio identification [5].

The diarization system is evaluated during the French ETAPE 2011("Evaluations en Traitement Automatique de la Parole") evaluation campaign [6] on TV and radio shows and obtained the best results among seven participants. The paper is organized as follows. In Section 2, the proposed speaker diarization system is presented. In Section 3, the experimental protocol and databases are described. Results and discussions are reported in Section 4. Conclusions and perspectives are given in Section 5.

Some of the programs used for this research were originally written by F. Bimbot, J. Černocký, A. El Hannani, G. Aversano and members of the French SYMPATEX project. The first author is financially supported by the ANR-SurfOnHertz project.

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

2. SPEAKER DIARIZATION SYSTEM BASED ON ALISP SEGMENTATION

The proposed speaker diarization system consists of the following steps:

a) ALISP-based audio sequencing and identification;
b) Voice activity detection;
c) Bayesian Information Criterion (BIC) segmentation;
d) BIC clustering;
e) Viterbi decoding; and
f) Normalized Cross Likelihood Ratio (NCLR) clustering.

Steps c) to f) are performed using the LIUM Speaker Diarization toolkit [7].

2.1. ALISP-based Audio Sequencing and Identification

The proposed system uses automatically acquired segmental units provided by ALISP tools to search for recurrent segments in TV and radio shows. The reference database is built from audio segments provided by annotated training and development databases (more details about databases are provided in section 3.1). These segments represent speech sentences, silence, noise, jingles, music and advertisements. Then ALISP transcriptions of reference segments are computed using HMMs (Hidden Markov Models) provided by the ALISP tools and compared to the transcriptions of the TV and radio shows stream using the Levenshtein distance [8].

An example of the output file provided by the ALISP module is shown in Figure 1. The ”Spk” label presents a recurrent speech sentence detected in the reference database, while the unknown label is relative to the signal part which was not detected in the reference database.

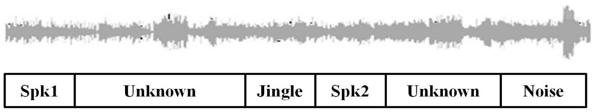

Fig. 1. Example of an output file provided by ALISP-based audio sequencing and identification

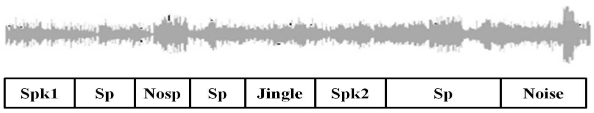

Fig. 2. Example of an output file provided by the voice activity detection system

ALISP method consists of two main modules: ALISP unit acquisition and segmentation, and approximate matching to find recurrent segments.

2.1.1. ALISP Model Acquisition and Segmentation

As explained in [2] [3] [4] and [5], the set of ALISP models is automatically acquired through parameterization, temporal decomposition, vector quantization, and Hidden Markov Modeling. This set of HMM ALISP models is used to transform a new incoming audio data into a sequence of ALISP symbols.

The parameterization of the audio data is done with Mel Frequency Cepstral Coefficients (MFCC), calculated on 20 ms windows, with a 10 ms shift. For each frame, a Hamming window is applied and a cepstral vector of dimension 15 is computed and appended with first order deltas.

After the parameterization step, temporal decomposition is used to obtain an initial segmentation of the audio data into quasi-stationary segments. This method was introduced originally by Atal [9] as a nonuniform sampling and interpolation procedure for efficient parameter coding. The detailed algorithm to compute interpolation functions can be found in [10].

The next step in the ALISP process is the unsupervised clustering procedure performed via Vector Quantization [11]. This method maps the P-dimensional vector of each segment provided by the temporal decomposition into a finite set of L vectors which define the number of ALISP units.

The final step is performed with the Hidden Markov Modeling procedure. The objective here is to train robust models of ALISP units on the basis of the initial segments.

Figure 3 shows an ALISP segmentation of an advertisement excerpt.

2.1.2. Approximate Matching Process of ALISP Units

Approximate string matching algorithms are a traditional area of study in computer science. With the huge increase of nucleotide and protein sequence data produced by various genome projects, fast string matching algorithms are developed. Our approximate string matching algorithm is based on the Basic Local Alignment Search Tool (BLAST) [12], widely used in bioinformatics. More details about this components can be found in [5].

2.2. Voice Activity Detection

The next step in the system is the voice activity detection. The goal is to remove the ”nonspeech” segments whose duration is above a predefined threshold.

Our voice activity detection system operates only on the portions of the signal labeled as ”unknown” by the ALISP-recognizer. It relies on a two-class detector, with Gaussian Mixture Model (GMM) trained on speech and non speech data.

The parameterization is done with Mel Frequency Cepstral Coefficients (MFCC), calculated on 20 ms windows, with a 10 ms shift. For each frame, a cepstral vector of di-

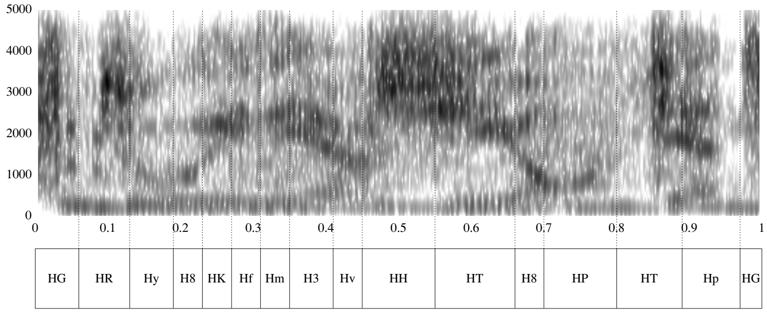

Fig. 3. Advertisement spectrogram with its ALISP segmentation, ”HG”,”HR”,”Hy”,etc. are the ALISP model names.

mension 12 is computed and appended with first and second order deltas and the Zero Cross Ratio.

A minimum duration of 0.5 s is defined for speech and nonspeech segments. In fact each class is modeled as a concatenation of 50 one-state HMM models.

An example of the output file provided by the voice activity detection module is shown in Figure 2. The ”Nosp” label is relative to a non speech segment, while the ”Sp” is relative to speech segment.

2.3. Bayesian Information Criterion - BIC Segmentation

The goal of speaker segmentation is to split an audio stream into homogeneous regions where only one speaker is present. After the voice activity detection step, only the portions of the signal labeled as ”Sp” are taken into consideration for the BIC segmentation.

This problem is considered as a model selection problem between two neighboring and overlapping audio segments as follows:

- a single model M for both segments $X_1 = x_1, ..., x_i$ and $X_2 = x_{i+1}, ..., x_N$,

- two different models M_1 and M_2 for the segments X1 and X2 respectively.

In practice, a metric distance is computed between the two hypothesis and an empirically set threshold is used to decide whether both segments come from the same speaker. The most common distance is based on BIC and its associated △BIC [13]. The two closest segments are merged at each iteration until $\triangle BIC > 0$.

2.4. Bayesian Information Criterion Clustering

Whereas the BIC segmentation operates on neighboring segments in order to detect whether or not they correspond to the same speaker, BIC clustering is performed to group together all the segments that belong to the same speaker. As for the segmentation process, at each iteration the closest clusters are merged until $\triangle BIC > 0$.

2.5. Viterbi Decoding

Viterbi decoding is performed to generate a new segmentation. Each cluster is modeled by a single-state HMM with an 8-component GMM. This process is necessary in order to refine the segment boundaries.

2.6. Normalized Cross Likelihood Ratio Clustering

Up to this level, the MFCC features are not normalized in order to use the information related to the background environment to detect speaker changes and ensure that each cluster contains one speaker. This step aims to avoid that several clusters represent the same speaker. Therefore, the background environment is removed, through feature normalization. Then, a hierarchical agglomerative clustering is realized over the last diarization.

An Universal Background Model (UBM) is built using training data, and the means of this model are adapted for each cluster to obtain a GMM model for each speaker. As for the BIC clustering, at each iteration, the two closest clusters are merged. The most common measure used in this step is the Normalized Cross Likelihood Ratio-NCLR [14]. The clustering stops when the NCLR measure is higher than a predefined threshold.

3. EXPERIMENTAL SETUP

As previously mentioned, the system was evaluated during the French ETAPE 2011 evaluation campaign. This campaign focused on TV and radio shows with various level of spontaneous speech and multiple speaker speech and did not target any particular type of shows.

Four tasks were considered in the ETAPE 2011 benchmark which are: multiple speaker detection, speaker diarization, lexical transcription, named entity detection.

3.1. Corpus

The ETAPE 2011 corpus consists of 13.5 hours of radio data and 29 hours of TV data. Table 1 summarizes the available data.

Note that the number of hours is reported in terms of recordings, not speech. It was measured that about 77% of the recording contains speech. Moreover, it was found that about 1.5 hours correspond to multiple speaker areas, which corresponds to about 7% of the time over all shows. This amount of overlapping speech makes the speaker diarization task more complicated.

genre	train	dev	test
TV news	7h30	1h35	1h35
TV debates	10h30	2h40	2h40
TV amusements	-	1h05	1h05
Radio shows	7h50	3h00	3h00
Total	25h50	8h20	8h20

Table 1. Duration of the training, development and test sets of ETAPE 2011 data.

3.2. Thresholds Setting

The proposed speaker diarization system contains four thresholds value which need to be fixed. These threshold are related to Levenshtein distance, BIC segmentation, BIC clustering and NCLR clustering.

Previous experiments were conducted in order to fix the Levenshtein distance threshold in the context of audio identification where the goal is to identify advertisements and songs in radio streams [5]. These experiments consist of computing the Levenshtein distance between ALISP transcriptions of the reference advertisements and their broadcasted occurrences in the radios and between ALISP transcriptions of the reference advertisements and data that does not contain advertisements. This study leads to a Levenshtein distance threshold of 0.55% [5].

In order to fix the other three thresholds, an automatically tuning by trying various combinations of thresholds was performed on the ETAPE development corpus. Each generated segmentation is scored against the reference segmentation and the thresholds that gave the lowest DER were used in the evaluation.

4. RESULTS AND COMPARISON

In order to evaluate the contributions of the ALISP-based module to the diarization results, a second experience was performed without that module. In Table 2 the DER values are reported for the baseline system (without the ALISP module) and the ALISP-based system.

Note that the ALISP-based module has improved the diarization results for all TV and radio shows. However, these improvements were not significant for all audio files. This is essentially related to the structure of the radio or TV show and whether this structure is the same each time the program is broadcasted.

Show name	Baseline	ALISP
BFMTV-BFMStory-175900	19.30	15.87
LCP-CaVousRegarde-235900	20.70	12.60
LCP-EntreLesLignes-192800-1	24.77	17.31
LCP-EntreLesLignes-192800-2	27.19	18.48
LCP-PilesEtFace-192800	28.42	19.76
LCP-TopQuestions-000400	35.46	29.55
LCP-TopQuestions-213800	15.87	2.44
TV8-LaPlaceDuVillage-201300	37.86	22.27
TV8-LaPlaceDuVillage-172800	35.82	20.40
EST2BC-FRE-FR-1000	14.55	13.75
EST2BC-FRE-FR-1750	39.41	22.93
EST2BC-FRE-FR-2152-1	41.83	27.34
EST2BC-FRE-FR-2152-2	29.91	23.93
EST2BC-FRE-FR-0910	8.73	8.26
EST2BC-FRE-FR-2004	21.13	15.48
ETAPE-2011 (whole data)	**24.73**	**16.23**

Table 2. Diarization Error Rate for the baseline and ALISP system.

Generally, the introduction of the ALISP module in the speaker diarization system has decreased the DER by 8.5%. Moreover, it is important to notice that the proposed ALISP-based speaker diarization system has obtained the best results in the ETAPE 2011 evaluation campaign among 7 participants, where the greatest DER value was 29.32% [15].

Related to the processing time, the system without the ALISP-based module runs at a speed of 10 seconds per minute on a 3.00GHz Intel Core 2 Duo 4GB RAM. When the ALISP-based module is added, the runtime increased to 40 seconds per minute.

5. CONCLUSIONS AND PERSPECTIVES

In this paper, a speaker diarization system using a data-driven segmentation process is proposed. A new module based on ALISP models is added before the segmentation and clustering process in order to identify similar audio segments between the same TV or radio show broadcasted on different dates. The system was evaluated during the ETAPE 2011 evaluation campaign and obtained a DER of 16.23%, which is the best result among all participants. We also demonstrate that the ALISP module in the speaker diarization system has decreased the DER by 8.5%.

Future work will be dedicated to extend this work to the visual context. The main idea is to train an audiovisual data driven model and exploit them in order to segment audiovisual document. We will also focus on improving the speaker diarization system by using the semantic information derived from an automatic speech recognition system.

6. REFERENCES

[1] X. Anguera, 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*, vol. 20, no. 2, pp. 356 –370, 2012.

[2] G. Chollet, J. Černocký, A. Constantinescu, S. Deligne, and F. Bimbot, *Towards ALISP: a proposal for Automatic Language Independent Speech Processing*, pp. 357–358, NATO ASI Series. Springer Verlag, 1999.

[3] A. El Hannani, D. Petrovska-Delacrétaz, B. Fauve, A. Mayoue, J. Mason, J.F. Bonastre, and G. Chollet, "Text independent speaker verification.," in *Guide to Biometric Reference Systems and Performance Evaluation*. Springer Verlag, 2009.

[4] G. Chollet, K. McTait, and D. Petrovska-Delacrétaz, "Data driven approaches to speech and language processing," *Lecture notes in computer science*, pp. 164–198, 2005.

[5] H. Khemiri, D. Petrovska-Delacrétaz, and G. Chollet, "A generic audio identification system for radio broadcast monitoring based on data-driven segmentation," in *IEEE International Symposium on Multimedia*, 2012.

[6] G. Gravier, G. Adda, Paulsson N., Carr M., Giraudel A., and Galibert O., "The etape corpus for the evaluation of speech-based tv content processing in the french language," in *International Conference on Language Resources and Evaluation*, 2012.

[7] S. Meignier and T. Merlin, "Lium spkdiarization: An open-source toolkit for diarization," in *CMU SPUD Workshop*, 2010.

[8] V. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Cybernetics and control theory*, vol. 10, pp. 707–710, 1966.

[9] A. Bishnu, "Efficient coding of lpc parameters by temporal decomposition," *ICASSP*, pp. 81–84, April 1983.

[10] F. Bimbot, "An evaluation of temporal decomposition," Tech. Rep., Acoustic Research Department AT&T Bell Labs, 1990.

[11] Y. Linde, A. Buzo, and R.M. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on Communication*, vol. 28, no. 1, pp. 84–95, 1980.

[12] S. F. Altschul, W. Gish, and W. Miller, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, pp. 403–410, 1990.

[13] S. S. Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *in Proc. DARPA Broadcast News Transcription and Understanding Workshop*, 1998, pp. 127–132.

[14] V. B. Le, O. Mella, and D. Fohr, "Speaker diarization using normalized cross likelihood ratio," in *Interspeech*, 2007, pp. 1869–1872.

[15] Evaluations en Traitement Automatique de la Parole, "http://www.afcp-parole.org/etape/workshop.html," 2012.