Application of Fischer semi discriminant analysis for speaker diarization in costa rican radio broadcasts

Aplicación del análisis semi discriminante de Fischer para la diarización de locutores en transmisiones de radio costarricenses

Roberto Sánchez-Cárdenas[1], Marvin Coto-Jiménez[2]

Sánchez-Cárdenas, R.; Coto-Jiménez, M. Application of Fischer semi discriminant analysis for speaker diarization in costa rican radio broadcasts. *Tecnología en Marcha*. Vol. 35, special issue. November, 2022. International Work Conference on Bioinspired Intelligence. Pág. 128-136.

https://doi.org/10.18845/tm.v35i8.6464

[1] University of Costa Rica. Costa Rica. E-mail: roberto.sanchezcardenas @ucr.ac.cr

[2] University of Costa Rica. Costa Rica. E-mail: marvin.coto@ucr.ac.cr
https://orcid.org/0000-0002-6833-9938

Keywords

Broadcasting; clustering; speaker diarization; speech technologies.

Abstract

Automatic segmentation and classification of audio streams is a challenging problem, with many applications, such as indexing multimedia digital libraries, information retrieving, and the building of speech corpus or spoken corpus) for particular languages and accents. Those corpus is a database of speech audio files and the corresponding text transcriptions. Among the several steps and tasks required for any of those applications, the speaker diarization is one of the most relevant, because it pretends to find boundaries in the audio recordings according to who speaks in each fragment. Speaker diarization can be performed in a supervised or unsupervised way and is commonly applied in audios consisting of pure speech. In this work, a first annotated dataset and analysis of speaker diarization for Costa Rican radio broadcasting is performed, using two approaches: a classic one based on k-means clustering, and the more recent Fischer Semi Discriminant. We chose publicly available radio broadcast and decided to compare those systems' applicability in the complete audio files, which also contains some segments of music and challenging acoustic conditions. Results show a dependency on the results according to the number of speakers in each broadcast, especially in the average cluster purity. The results also show the necessity of further exploration and combining with other classification and segmentation algorithms to better extract useful information from the dataset and allow further development of speech corpus.

Palabras clave

Radiodifusión; agrupación; registro de locutores; tecnologías del habla.

Resumen

La segmentación y clasificación automática de transmisiones de audio es un problema desafiante, con muchas aplicaciones, como la indexación de bibliotecas digitales multimedia, la recuperación de información y la construcción de corpus de voz (o corpus hablado) para idiomas y acentos particulares. Ese corpus es una base de datos de archivos de audio de voz y las transcripciones de texto correspondientes. Entre los varios pasos y tareas requeridos para cualquiera de esas aplicaciones, la diarización del hablante es una de las más relevantes, porque pretende encontrar límites en las grabaciones de audio según quién habla en cada fragmento. La diarización del hablante se puede realizar de forma supervisada o no supervisada y se aplica comúnmente en audios que consisten en habla pura. En este trabajo, se realiza un primer conjunto de datos anotados y análisis de la diarización de locutores para la radiodifusión de Costa Rica, utilizando dos enfoques: uno clásico basado en la agrupación de k-medias y el más reciente Fischer Semi Discriminant. Elegimos la transmisión de radio disponible públicamente y decidimos comparar la aplicabilidad de esos sistemas en los archivos de audio completos, que también contienen algunos segmentos de música y condiciones acústicas desafiantes. Los resultados muestran una dependencia de los resultados de acuerdo con el número de hablantes en cada transmisión, especialmente en la pureza promedio del clúster. Los resultados también muestran la necesidad de una mayor exploración y combinación con otros algoritmos de clasificación y segmentación para extraer mejor información útil del conjunto de datos y permitir un mayor desarrollo del corpus del habla.

Introduction

Speaker diarization is a process that involves the segmentation and clustering of audio recordings, to determine when each of the participants in the recording speaks. It can also be described as the process of partitioning an input audio stream into homogeneous segments according to speaker identity [1].

Due to the development and availability of massive audio-visual data during the past few years, the efficient management of audio content is becoming inevitable [2]. There is a need to implement techniques to process the video and audio data automatically, and speaker diarization is one of such leading techniques that can allow the clustering and characterization of audio information in terms of "who speak when."

The process is also useful for many speech processing technologies which assume the presence of only one speaker, such as automatic speaker identification. For theses processes, speaker diarization can be an essential front end where the single-speaker assumption cannot be considered [3].

The process involved in performing speaker diarization can be described as the clustering of segments of the same acoustic nature. This has been implemented to detect segments of the same speaker, but theoretically it can be also be applied to other kind of sounds. For specific cases and conditions, it is expected that the different algorithms available perform distinctly.

In radio broadcast signals, diarization can be part of a system that detects audio parts that contain speech, music, silence and other types of sounds. Then, each part detected can be processed by speech recognizers, language recognizers, singer recognizers, song recognizers, etc. [4].

The speech corpus developed for this paper consists of the annotation of the periods of time where a labeled speaker participated in the audio stream. For the building of speech corpus that can be part of speech recognition systems or speech synthesis systems, the automatic characterization of broadcasts, in terms of longitude of speech (or music) segments, can be of great interest. This is due to the difficulties and long processes it takes to build a corpus using studio recordings.

This is the case of the present paper, where the development of speech technologies suitable for Costa Rican Spanish requires the building of large corpus of speech recordings and its corresponding transcription and labeling. The use of radio broadcasts for such research has been widely explored for several languages [5–8].

The rest of this paper is organized as follows: Section 2 provides a brief overview about Related work and the Fisher Semi-Discriminant Analysis. Section 3 describes the Experimental Setup. Section 4 presents the Results and discussion, and finally, Section 5 presents the conclusions.

Background

Related work

The building of corpus for research in speech technologies using radio broadcast have been presented in [9], for an Italian Broadcast News Corpus. There are many tasks required to segment, annotate and cluster speech from a such source of audio data. Speaker diarization is one of the first and most important of those tasks.

In [1] the diarization of broadcast recordings is based on an audio partitioner, which provides high cluster purity (99% for the best combination of features). The experiments were conducted on US English and French data from broadcast news.

More recently, the Fisher Semi-Discriminant Analysis (FLSD), developed and presented in [10] have been applied successfully for this task. In its first report, using a large corpus of 70 debate with 190 participants, using the Canal 9 Database. A refinement of the proposal was presented in [11], using different longitude of the segments in the process of clustering, providing relative improvement than the previous one.

FLSD has also been combined with visual features (such as face expression) to improve results [12]. In our work, we build an initial database of Costa Rican radio broadcast for the experimentation with the FLSD. In our case, only the audio recordings are available, so the combination with visual features is not possible. We chose to perform this first experimentation with the raw podcast as the source recordings for diarization. This also means short segments of music, advertisements and other factors are present, which represents new challenges for the algorithm.

Fisher Semi-Discriminant Analysis

FLSD is an extension of the Fisher linear discriminant analysis (FLD) method, which is used in classification in a similar way to other dimensionality reduction techniques (e.g. Principal Component Analysis), but with the clustering guided by the information from an existing mapping between linear combination of features, x with a corresponding set of classes ck [12].

For this purpose, a set of scatter matrices are calculated from between class and within class vectors. The purpose of FLD is to perform a reduction where the class-means are well separated, measured relative to the data assigned to a particular class [13].

The advantages of FLD relies on the employ of information from the classes, but this advantage cannot be considered in unsupervised cases, such as speaker diarization. For this reason, FLSD, presented in [10] only requires a set of features vectors that belong to the same class, given any feature vector.

In temporal data such as an audio recording, it is expected that given any sample of the audio, the neighbor samples most likely belong to the same speaker or music segment. With this information it is possible to estimate the scatter matrices required in FLSD closer to those of the case where the original classes are known. Further details of this technique can be found in [12, 10]. In our work, we use the implementation presented in [14].

Experimental setup

Database

In order to evaluate the performance of the Fisher Semi-Discriminant Analysis in a Spanish speaking database, a new corpus was developed. This corpus is based on the Costa Rican radio program obtained from Radio Universidad, a broadcast from the University of Costa Rica. The program used for this corpus is named *Desayunos* (Breakfasts), it has different hosts, guests and ads every day, which makes it an ideal program for the evaluation of the algorithm.

A total of ten complete programs were used and divided in half to obtain more accurate results. Every program was annotated in a tab-separated value format.

The test runs where made on raw audio, which contained speakers, silence, music and other sounds from the ads. Most programs were recorded using an internet communication stream, which caused quality losses in the audio in small periods.

Evaluation

As stated in [10], a set of two common metrics is provided by FLSD method: Average Speaker Purity (ASP) and Average Cluster Purity (ACP). Both measurements require a ground truth of annotated speaker turns in the audio files.

The cluster purity measures the frequency of the most common speaker into each cluster. The higher the ACP means higher information of a unique speaker within each cluster. Speaker purity is based on the frequency of the most common detected speaker within each speaker class. The higher the ASP means a better coincidence of the detected speaker turns.

Results and Discussion

Two main tests were made to the data; with a known and an unknown number of speakers in a specific program. This was made in order to prove the capabilities of the software to be used in supervised and unsupervised diarization. Due to the annotations a percentage of cluster and speaker purity is obtained. Every program was tested on the basic speaker diarization algorithm and FLSD.

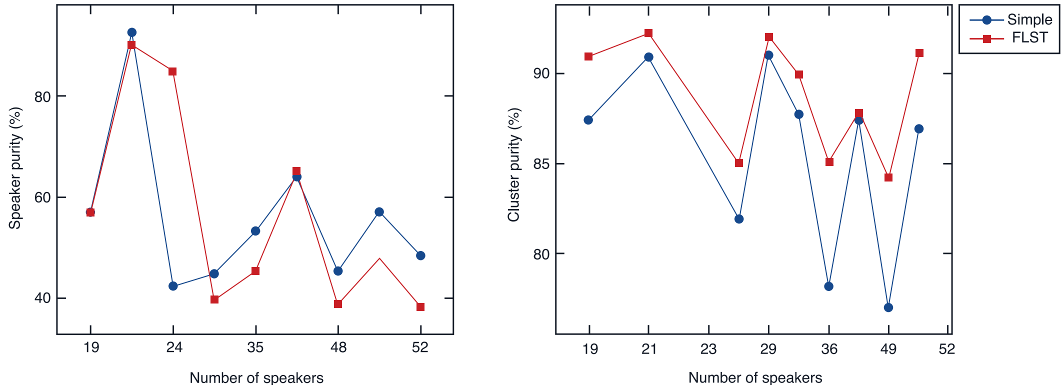
Table 1 presents the result from test runs where the number of speakers wasn't specified beforehand. The results of the base system (based on K-means) and FLSD are shown side by side in order to compare them. The ACP variate in a range from 37.6% to 98% using the base system, while FLSD has a range from 49.7% to 100%. FLSD improved the diarization up to 16.7% and had no significant worsening. Overall, FLSD improved cluster purity by 3.2%, where there is an unknown number of speakers. The standard deviation was reduced by 14.34 using FLSD, which means it associates speakers more accurately.

As for speaker purity, the base system had a range from 26.9% to 98.5%, while FLSD varies from 54.6% to 99.9%. Overall, FLSD improved speaker purity by 8,4% and had a standard deviation of 13.02 compared to 19.26 using the base system.

It can be seen in Figure 1, that with an unknown number of speakers, the speaker purity is almost constant. Cluster purity tends to decrease significantlyas there are more speakers and it is noticeable that both FLSD and simple method perform similarly.

Table 1. Cluster and speaker purity. Unknown number of speakers.

	Base system		FLSD	
Program	ACP (%)	ASP (%)	ACP (%)	ASP (%)
1	37.6	72.7	54.3	
2	68.1	73.2	79.4	98.0
3	86.5	97.6	87.1	97.7
4	86.0	97.9	86.1	98.0
5	48.0	88.1	50.1	82.5
6	43.0	84.6	49.7	93.1
7	65.0	99.6	64.9	99.6
8	62.1	99.0	62.0	99.2
9	76.2	82.7	78.6	99.9
10	82.0	62.0	85.2	99.9
11	75.2	92.7	77.7	95.4
12	74.9	97.7	74.8	97.2
13	66.4	79.5	66.8	97.7
14	59.6	94.4	70.3	97.3
15	65.4	97.6	65.4	97.7
16	74.1	98.5	74.3	98.9
17	81.8	95.5	82.9	95.5
18	74.2	87.5	79.6	88.0
19	98.0	26.9	99.2	57.8
20	97.5	48.4	100	54.6
MEAN & SD	71.08 (16.14)	83.80 (19.26)	74.28 (14.34)	92.20 (13.02)



(a) Speaker purity as a function of number of speakers

(b) Cluster purity as a function of number of speakers

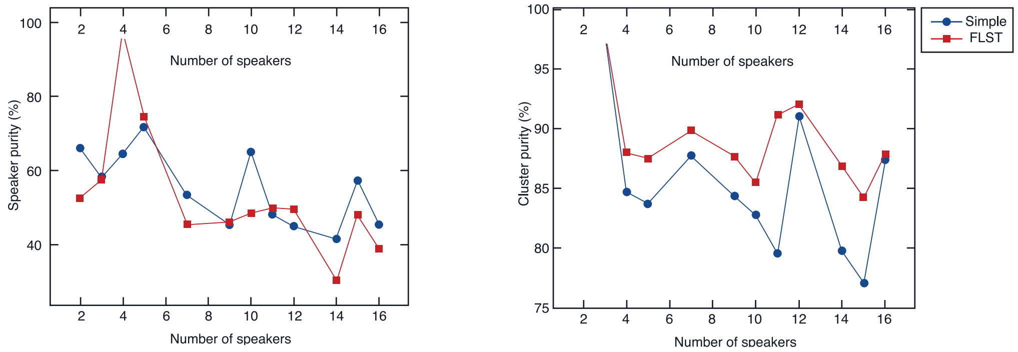
Figure 1. Unknown number of speakers ACP and ASP results.

as there are more speakers and it is noticeable that both FLSD and simple method perform similarly.

Supervised diarization results are presented in Table 2. With this method cluster purity varies in a range from 70% to 98% in the base system, while FLSD varies from 82.6% to 99.2%. On average the results were improved by 5,06%. ACP using FLSD had an average of 90.17% and a standard deviation of 4.63. There's an overall improvement using the FLSD method compared with the base method.

Table 2. Cluster and speaker purity. Fixed number of speakers.

Program	Base system		FLSD	
	ACP (%)	ASP (%)	ACP (%)	ASP (%)
1	87.4	45.3	87.8	38.7
2	91.0	44.8	92.0	39.5
3	86.9	48.4	91.1	38.3
4	87.7	53.3	89.8	45.4
5	79.5	48.1	87.5	49.8
6	70.0	57.2	84.2	48.0
7	90.9	92.5	92.2	90.1
8	76.8	80.0	82.6	80.6
9	81.9	42.3	85.0	85.0
10	87.4	57.0	90.9	56.9
11	84.3	45.3	87.6	46.0
12	84.7	46.2	87.1	33.4
13	79.7	41.5	86.8	30.2
14	82.7	64.9	85.4	48.3
15	82.6	71.0	94.8	74.7
16	82.5	72.4	92.3	74.2
17	83.4	64.4	97.2	97.2
18	80.2	56.8	92.4	63.5
19	98.0	58.1	99.2	57.8
20	97.5	66.0	97.5	52.4
MEAN & SD	85.10 (5.90)	57.58 (13.71)	90.17 (4.63)	57.5 (19.80)

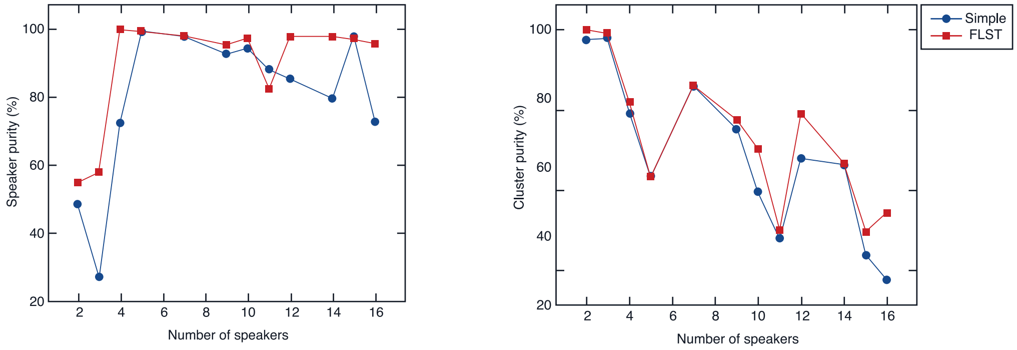


(a) Speaker purity as a function of number of speakers

(b) Cluster purity as a function of number of speakers

Figure 2. Fixed number of speakers ACP and ASP results.

In Figure 2 its noticeable that when using a fixed number of speakers, the speaker purity tend to decrease as there are more speakers and most results are below 60% and over 30%. Cluster purity with FLSD method doesn't seem to present a significant decrease as the number of speakers increase. The simple method appears to decrease slightly more than FLSD.



(a) Speaker purity as a function of number of speakers

(b) Cluster purity as a function of number of speakers

Figura 3. Speaker purity with a greater number of segments.

As shown in Figure 3, in both the base system and FLSD the speaker purity decreases with a greater number of segments. Cluster purity does not present a significant pattern of decrease with a greater number of segments. The base system and FLSD have similar behavior.

Conclusions

In this paper, we performed an analysis of the FLSD algorithm to experimentally validate its applicability in Costa Rican radio broadcast data, where the different conditions of the audio, and the presence of music segments and advertisements are challenging.

The FLSD diarization method had good results when there was a fixed number of speakers. On average, it matched the correct speaker 90.17% of the time. But the same method didn't perform similarly well with an unknown number of speakers as it only matched the speaker correctly 74.28% of the time. The outcomes show that with fewer participants, the result is better for both supervised and unsupervised diarization.

Cluster purity with a fixed number of speakers could have gotten worse due to the participants that appeared in ads for only short periods of time since the algorithm does not know beforehand how much participation a speaker had and it requires the speakers to appear more in order to identify them and associate them correctly. Also, some speakers had background music which makes it harder to correctly identify the speakers.

The speaker purity results are not as good as the cluster purity. It has to be taken into account that the database used is based on raw audio from a radio broadcast, which had music in different segments and signal losses due to the digital audio stream.

References

[1] Barras, Claude, et al. “Multistage speaker diarization of broadcast news.” IEEE Transactions on Audio, Speech, and Language Processing 14.5 (2006): 1505-1512.

[2] Vavrek, Jozef, et al. “Classification of broadcast news audio data employing binary decision architecture.” Computing and Informatics 36.4 (2017): 857-886.

[3] García-Romero, Daniel, et al. “Speaker diarization using deep neural network embeddings.” 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017.

[4] Theodorou, Theodoros, Iosif Mporas, and Nikos Fakotakis. “An overview of automatic audio segmentation.” International Journal of Information Technology and Computer Science (IJITCS) 6.11 (2014): 1.

[5] Pleva, Matu´s, and Jozef Juh´ar. “TUKE-BNews-SK: Slovak Broadcast News Corpus Construction and Evaluation.” LREC. 2014.

[6] Yilmaz, Emre, et al. “A longitudinal bilingual Frisian-Dutch radio broadcast database designed for code-switching research.” (2016).

[7] Zgank, Andrej, Ana Zwitter Vitez, and Darinka Verdonik. “The Slovene BNSI Broadcast News database and reference speech corpus GOS: Towards the uniform guidelines for future work.” LREC. 2014.

[8] Nouza, Jan, Jindrich Zdansky, and Petr Cerva. “System for automatic collection, annotation and indexing of Czech broadcast speech with full-text search.” MELE – CON 2010-2010 15th IEEE Mediterranean Electrotechnical Conference, 2010.

[9] Federico, Marcello, Giordani, Dimitri and Coletti Paolo. “Development And Eval – uation Of An Italian Broadcast News Corpus.” European Language Resources Association (ELRA). 2000.

[10] Giannakopoulos, Theodoros, and Sergios Petridis. “Fisher linear semi-discriminant analysis for speaker diarization.” IEEE transactions on audio, speech, and language processing 20.7 (2012): 1913-1922.

[11] Montazzolli, Sergio, Andre Adami, and Dante Barone. “An extension to Fisher Linear Semi-Discriminant analysis for Speaker Diarization.” 2014 International Telecommunications Symposium (ITS). IEEE, 2014.

[12] Sarafianos, Nikolaos, Theodoros Giannakopoulos, and Sergios Petridis. “Audiovisual speaker diarization using fisher linear semi-discriminant analysis.” Multimedia Tools and Applications 75.1 (2016): 115-130.

[13] Welling, Max. “Fisher linear discriminant analysis”. Department of computer science, University of Toronto. Technical Report, 2005.

[14] Giannakopoulos, Theodoros. “pyaudioanalysis: An open-source python library for audio signal analysis.” PloS one 10.12 (2015): e0144610.