Smart sound sensor to detect the number of people in a room

Sami Boutamine, Istrate Dan, Jérôme Boudy

Smart sound sensor to detect the number of people in a room

Sami Boutamine[1,2], Dan Istrate[1], Jérôme Boudy[2]
[1]Sorbonne Université, Université de Technologies de Compiègne, BMBI UMR7338, France
[2]Télécom Sud Paris, SAMOVAR-ARMEDIA UMR 5157 Evry, France

***Abstract*—In order to allow older adults to active and healthy ageing with comfort and security we have already developed a smart audio sensor being able to recognize everyday life sounds in order to detect activities of daily living (ADL) and distress situations. In this paper, we propose to add a new functionality by analyzing the speech flow in order to detect the number of person in a room. The proposed algorithms are based on speaker diarization methods. This information is useful in order to better detect activities of daily life but also to know when the person is home alone. This functionality can also offer more comfort through light, heating and air conditioning adaptation to the number of persons.**

I. INTRODUCTION

Nowadays, technologies at the service of intelligent spaces are constantly developing and interacting with everyday life objects. They exploit video, audio signals and environmental data to locate people, to recognize their gestures, to interact with the people in order to offer comfort and to help people with disabilities. An important domain of study is the Ambient Assisting Living (*AAL*) which try using the new technologies to compensate the disabilities related to the age (visual, hearing, cognitive).

We have already developed a smart audio sensor[1] allowing the detection of distress situations but also of people activity through sound environment analysis. This proposed sensor analyse the sound environment and is able using different techniques to recognize 18 sound classes in a continuous audio flow.

In this paper, we propose a new functionality to the already presented sensor by adding the possibility to detect the number of persons present in a room through speech analysis. In fact, knowing the number of persons help to know if the elderly people is home alone and to adapt the distress situations identification. Also, knowing the presence of other persons allow to follow the intervention at home of different services (catering, cleaning services,). Otherwise, combining the speech with sound analysis allow a better activity daily life identification in order to allow an Active and Healthy Ageing (*AHA*) of young older adults.

Additionally, this new functionality allow also offering comfort and energy consumption reduction by adapting the light quantity, the heating or the air conditioning systems.

This work is a part of the multi-partner national project CoCAPs (FUI type). In this article, we present the work and results obtained from the detection of the number of speakers where several tools and methods of sound processing are used including speaker diarization. This paper is organized as follows, we present first the speaker diarization system, then, the developed application (detection of the number of speakers) will be presented by detailing the implementation of the application of sound detection system (speech/no speech) by using some functionalities offered by the LIUM_SpkDiarization toolkit. Finally, we present and discuss the first test results of the developed application.

II. SPEAKER DIARIZATION SYSTEM

Speaker diarization is to determine "who spoke when?" in an audio record that contain speech, music or noise segments. The signal audio is splitted into homogeneous speech segments, according to the speaker identity with no prior knowledge about it.

The principal modules of diarization system are composed of parametrization, speech activity detection or voice activity detection, speaker segmentation, speaker clustering and re-segmentation. Already available tools are LIUM_SpkDiarization [2], audioSeg, DiarTK, and SHoUT for this purpose.

A. *Parametrization*

Mel-Frequency Cepstral Coefficients (*MFCC*), Perceptual Linear Prediction coefficients (*PLP*) and Linear Frequency Cepstral Coefficients (*LFCC*) are the most common features used to extract the most important information from the signal, sometimes used with their first and/or second derivatives.

Ideal features should have some important properties:

- They have to emphasize the difference between classes (class separability).
- Their intra-class variances must be minimal.
- They have to be robust to noise trouble, conserving the class separability as far as possible.
- The total number of detected features (i.e. features density) should be sufficiently large to reflect the frames content in a compact form.
- A high correlation between features should be avoided, as much as possible.

The *MFCCs* are widely used in automatic speech and speaker recognition, and used for this current work.

B. *Speech activity detection*

The speech activity detection plays a very important role in the whole diarization process, it is performed to separate speech, no-speech and silent frames.The output of other sub task shighly depends on the precision of this task. The goal is to keep only relevant information for speech modeling and to reduce the amount of computation

needed for the following treatments. Many approaches allow the detection of speech, segmentation using Hidden Markov Models (*HMMs*) [3], Deep Neural Networks (*DNN*) [4]. In this work, we used an algorithm based on the Wavelet Transform [5].

C. *Speaker segmentation*

The aim of the speaker segmentation is to find speaker change points in a given speech signal using the symmetric Kullback Leibler distance (*KL2*), the generalized likelihood ratio (*GLR*) or the Bayesian information criterion (*BIC*) distance computed using Gaussians with full covariance matrices. In such systems, the speech signals are windowed for a short duration of 25 – 30ms.

In such systems like LIUM_SpkDiarization toolkit [6], the segmentation is done in two steps. A first pass on the signal is performed to detect breaks (changes in speakers), using the *GLR* (Generalized Likelihood Ratio) measure defined in "Eq. (1)". The *GLR* measure introduced by Gish [7] is a likelihood ratio between two hypotheses H_0 and H_1.

H_0: The two sequences x_i and x_i are produced by the same speaker x, which, in this case, the model $M\ (\mu, \Sigma)$ corresponding to $x = x_i \cup x_j$ would allow a better representation of x_i and x_j.

H_1: The two sequences x_i and x_i are produced by two different speakers, which case the two models $M_i\ (\mu_i, \Sigma_i)$ and $M_j\ (\mu_j, \Sigma_j)$ would be better suited to represent x_i and x_j.

The likelihood test is thus formulated by the ratio of the two hypotheses:

$$GLR(x_i, x_j) = \frac{L(x, M(\mu, \Sigma))}{L(x_i, M_i(\mu_i, \Sigma_i))L(x_j, M_j(\mu_j, \Sigma_j))} \quad (1)$$

Where $L(x, M(\mu, \Sigma))$ corresponds to the likelihood of the sequence $x = x_i \cup x_i$ given the model $M\ (\mu, \Sigma)$, and $L(x_i, M_i(\mu_i, \Sigma_i))L\ (x_j, M_j(\mu_j, \Sigma_j))$ the likelihood that the x_i and x_j were produced by two different speakers.

A second pass, allows to refine the segmentation obtained during the first pass by grouping consecutive segments that maximize a likelihood score using a discriminating measure *BIC* (Bayesian Information Criterion) defined in "Eq.(2)" below. *BIC* is a metric highly appreciated for its simplicity and efficiency.

$$BIC = -2\ln(L) + k\ln(N) \quad (2)$$

With L the likelihood of the estimated model, N the number of observations in the sample and k the number of free parameters of the model.

D. *Speaker Clustering*

Clustering is done using an unsupervised method called hierarchical agglomerative clustering (*HAC*). The goal of speaker clustering is to associate segments from an identical speaker together. Speaker clustering ideally produces one cluster for each speaker with all segments from a given speaker in a single cluster.

The initial set of clusters is composed of one segment per cluster. Each cluster is modeled by a Gaussian with a full covariance matrix. ΔBIC measure is employed to select the candidate clusters to group as well as to stop the merging process. The two closest clusters i and j are merged at each iteration until $\Delta BIC_{i,j} > 0$. ΔBIC is defined in "Eq. (3)" [6].

$$\Delta BIC_{i,j} = \frac{n_i+n_j}{2}\log|\Sigma| - \frac{n_i}{2}\log|\Sigma_i| - \frac{n_j}{2}\log|\Sigma_j| - \lambda P \quad (3)$$

$$P = \frac{1}{2}\left(d + \frac{d\,(d+1)}{2}\right) + \log(n_i + n_j) \quad (4)$$

Where $|\Sigma_i|$, $|\Sigma_i|$ and $|\Sigma|$ are the determinants of Gaussians associated to the clusters i, j and $i + j$. λ is a parameter to set up. The penalty factor P "Eq. (4)" depends on d, the dimension of the features, as well as on n_i and n_i, referring to the total length of cluster i and cluster j respectively.

This penalty factor only takes the length of the two candidate clusters into account whereas the standard factor uses the length of the whole data.

E. *Re-segmentation*

Re-segmentation is the final stage of the process, in which the rough boundaries of diarization systems that rely on segment clustering of an initial uniform segmentation are refined based on a frame-level. The most common approach is a Viterbi re-segmentation with *MFCC* features[2].

III. APPLICATION

The purpose of the application developed is the detection of the number of speakers in a room, a meeting room or an office, the development was realized on the Raspberry *Pi3 Model B* board.

In order to achieve our objective, we used the speaker diarization technique, whose goal is to segment the audio signal into small homogeneous regions containing only speech and which belongs to one and only one speaker. The number of speakers corresponds to the number of segments group obtained from each speaker.

A. *Application architecture*

The basic architecture of the application is shown in "Fig.1".

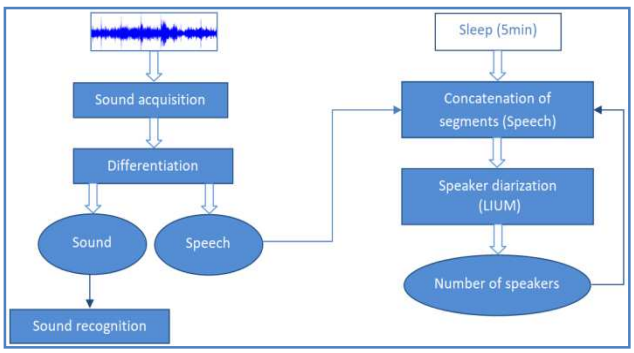

Figure 1. Application architecture.

It is an application that works in parallel, using two threads. In the first thread an application proceeds in a continuous way to capture the sound signal and differentiate it between the speech and the other sounds of everyday life whose objective is the recording of the segments containing only speech.

In the second thread, the program gets back and groups the segments of saved speech for a predefined duration in order to move to the segmentation phase using the LIUM_SpkDiarization toolkit. The output of this phase is represented by a file containing groups of segments with the speech of each speaker. The number of groups corresponds to the number of people detected automatically by speaker diarization.

B. Sound detection system (speech/no speech)

After analyzing the first results of the toolkit LIUM_SpkDiarization applied to a complete audio signal containing speech and human sounds of everyday life, it has been noticed that sounds, which are different from the speech, have a negative influence on the results.

Indeed a sound detection application was applied before using the toolkit LIUM_SpkDiarization, whose purpose is to filter the audio signal by removing all other sounds different from the speech.

The sound detection aims to detect, and separate speech events from other sound events in the continuous audio flow.

The classification of sound/speech is based on two Gaussians Mixtures Models (*GMM*). One class for speech and another one for the sounds of everyday life.

C. LIUM_SpkDiarization toolkit

LIUM_SpkDiarization is an open source diarization toolkit designed for extraction of speaker identity from audio records with no information before about the analysed data (number of speakers, etc.). *LIUM* could identify speaker's speech segments at excellent level.

LIUM_SpkDiarizationwas developed by *LIUM* (Computer Science Laboratory of Le Mans) for the French *ESTER2* evaluation campaign [6].

LIUM_SpkDiarization comprises a full set of tools to create a complete system for speaker diarization, going from the audio signal to speaker clustering based on the CLR/NCLR metrics. These tools include MFCC computation, speech/no-speech detection, and speaker diarization methods [8].

IV. EXPERIMENTS AND RESULTS

The first tests done to evaluate the developed application "detection of the number of persons" are performed in offices, taking into account the number of speakers and the duration of the audio segments to be processed.

In these first tests, the system was initially applied to segments containing only continuous speech and without noise in order to know the performance of the algorithm in in better conditions.

The properties of the processed audio signal are given in "Tab. I".

TABLE I. SPEECH CORPUS

Language	French
Sampling rate	16 Khz
N° of channels	1 (16-bits mono channel)
Speech domain	Conversational speech

The results obtained are shown by the curves in "Fig. 3" and in "Tab. II", the numbers represent the percentages of the algorithm performance in relation to the number of speakers and the durations of the audio segments.

The results represent the performance of the application "detection of the number of persons", applying the sound classification speech / no-speech to remove any different sounds of speech, followed by the speaker diarization tool (LIUM_SpkDiarization), the suppression of sound has improved the results.

The segmentation performance is calculated according to the following formula:

$$\text{Diarization\_performance} = \frac{\text{Number of files correctly identified}}{\textit{total number of files}}$$

The tests were done on continuous speaking times of 3/5/10 minutes after removing all the noises. This test should highlight the optimal duration for the algorithm to have enough data for models training.

TABLE II. TEST RESULTS IN TERMS OF DIARIZATION PERFORMANCE

	3min	**5min**	**10min**
1 Speaker	70%	90%	90%
2 Speakers	50%	70%	90%
3 Speakers	90%	70%	80%
Total	70%	76.66%	86.66%
Global performance	**77.77%**		

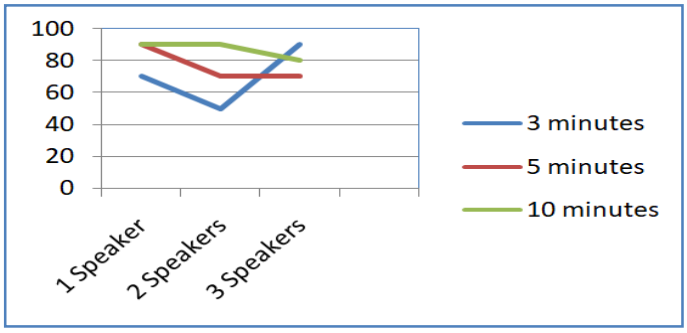

Figure 2. Diarization performance in function of duration.

From the first results, we notice that the algorithm performance is better for the longer durations (10 minutes).

Compared to the energy control application a window of 5-10 minutes is optimal because the current motion sensors use a delay of 15-20 minutes.

We present in "Tab. III", an example where the algorithm does not make an error and in "Tab. IV", an example with an error of a (+/-) 1 speaker.

S1, S2 and S3 correspond to the speaker label found automatically by the algorithm, M and F mean respectively Male and Female, the duration represent the time that each speaker has spoken.

TABLE III. EXAMPLE TESTS (10 MINUTES, 2 WOMEN, 1 MAN)

	S1(M)	**S2(F)**	**S3(F)**
Speaker1(M)	4min-23sec	X	X
Speaker2(F)	X	2min-02sec	X
Speaker3(F)	X	X	2min-16sec

TABLE IV. EXAMPLE TESTS (5 MINUTES, 3 WOMEN)

	S1(F)	**S2(F)**	**S3(F)**	**S4(F)**
Speaker1(F)	1min-04sec	X	X	X
Speaker2(F)	X	53sec	X	X
Speaker3(F)	X	X	1min-48sec	27sec

We notice in table IV that the third speaker has been identified by the system as two speakers (S3 and S4) and this because the system is sensitive to the change of prosody, works are underway to solve this problem.

The algorithm can make the difference between man and woman with a global error rate equal to 0%, and this represents very important information in our application, especially for the recognition of person in a room.

The evaluation of the application is underway, notably for a larger number of speakers and in different environments such as living rooms and meeting rooms, the aim of which is to improve the results obtained.

V. CONCLUSION AND FUTURE WORK

This paper proposes a new functionality to an existing smart audio sensor being able to recognize everyday life sound for *ADL* and distress detection. The new functionality allow the estimation of the number of speaker in the speech flow, information useful for *ADL* detection but also for adaptation of distress detection system when the person is home alone.

The results of this work corresponds to the tests of the speaker number detection application based on both speaker diarization (LIUM_SpkDiarization tool) and the application of the sound classification (Speech / Sound), by removing any sounds different from speech before applying the speaker diarization tool, whose goal is to improve the results obtained by LIUM_SpkDiarization.

For the future, the results will be improved by working on the sensor location optimization and the audio signal filtering. In addition, works are in progress to have better results with segments containing both speech and noise.

The use of speaker recognition methods is potentially considered as to improve the remote monitoring for elderly people in a room.

ACKNOWLEDGMENTS

The authors would like to thank BPI France, the Regional Councils of Limousin and Rhône-Alpes associated with the ERDF program, the departmental council of Isère, and the Bourges agglomeration community for their financial support to the project CoCAPs.

The CoCAPs project, from FUI N ° 20, is also supported by the poles of competitiveness S2E2 and Minalogic.

REFERENCES

[1] M. Robin, D. Istrate and J. Boudy, "Remote monitoring, distress detection by slightest invasive systems: Sound recognition based on hierarchical i-vectors," 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Jeju - Korea, 2017, pp. 2744-2748, doi: 10.1109/EMBC.2017.8037425

[2] S. Meigner and T. Merlin, "An Open Source Toolkit ForDiarization Sylvain Meignier, Teva Merlin LIUM – Université du Mans, France."

[3] RENEVEY P. & DRYGAJLO A. (2001). Entropy based voice activity detection in very noisy conditions. p. 1887–1890.

[4] RYANT N., LIBERMAN M. & YUAN J. (2013). Speech activity detection on youtube using deep neural networks. In INTERSPEECH, p. 728–731.

[5] D. Istrate, E. Castelli, M. Vacher, L. Besacier and J.-F. Serignat, Medical Telemonitoring System Based on Sound Detection and Classification, IEEE Transactions on Information Technology in Biomedicine, vol. 10, no. 2, Avril 2006.

[6] http://www-lium.univ-lemans.fr/diarization/doku.php

[7] H. Gish, M. H. Siu, and R. Rohlicek, Segregation of speakers for speech recognition and speaker identification, Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, Canada, vol. 2, pp. 873-876, 1991

[8] Eva KIKTOVA, Jozef JUHAR, Comparison of Diarization Tools for Building Speaker Database, vol.13, no.4, (2015)

[9] 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.