

# Impact of Multi-Classifer Fusion on Target Speaker Detection in Audio Streams

O. Kenai

Departments of Performing Arts and Audiovisual Studies, Higher Institute of Performing Arts Professions  
ISMAS / HIPAP  
Wassilakan17@gmail.com / kenai.ouassila@ismas.dz

DOI: 10.5281/zenodo.17618334

**Abstract**— This article discusses robust system by multi-classifier fusion approach used in target Speaker Detection (SD) systems to improve their performance. Single classifiers may introduce significant performance degradation in the performance. To overcome this problem, we propose in this work to apply the fusion of multi-classifiers Hierarchical Ascending Clustering (HAC), Gaussian Mixture Model (GMM) and Support Vector Machine (SVM) on an architecture based on Activity Detection Voice (VAD) in order to reduce errors of speakers' detection. A comparative investigation was conducted between individual classifiers and their fusion; and for the evaluation task, the three classifiers and their fusion were tested on telephonic conversations extracted from the NIST-2005 corpus. The results of experiments have shown that the applied multi-classifier fusion on this architecture has considerably enhanced the performances of target SD system, comparing to the applied each classifier. The results show a Speaker Detection Rate (SDR) of 99.18% with the fusion approach, compared to HAC (85.98%), GMM (86.68%), and SVM (97.67%).

**Keywords**— *Speaker Detection (SD), Voice Activity Detection (VAD), Hierarchical Ascending Clustering (HAC), Gaussian Mixture Model (GMM), Support Vector Machine (SVM), Multi-classifier Fusion, Majority Voting.*

## 1 Introduction

Searching for specific content in audio, audiovisual, and multimedia documents is an essential task. Various approaches and techniques are employed to extract specific information from this vast collection of documents, such as speaker identities, ages, genders, or other attributes, depending on the application.

In the case of audio documents, pre-defined information is often used for classification purposes. A key element is speaker identity, which is leveraged to locate information related to a target speaker. Consequently, numerous Speaker Detection (SD) systems have been developed to fulfill this task (Reynolds 2004), (Vesnicer 2008). Many of these systems utilize classifiers such as Hierarchical Ascending Clustering (HAC), Gaussian Mixture Models (GMM), or Support Vector Machines (SVM), among others, to identify target speakers in audio content (Anguera 2012), (Kenai 2018), (Shon 2019), (Kenai 2019a).

SD is a task linked to speech processing resulting from the increase in the number of multimedia documents that need to be properly segmented, clustered, detected archived. In that sense, the goal of SD is to segment an audio stream in homogeneous segments containing the

voice of only one speaker (also called speaker change detection process) and to associate the resulting segments by matching those belonging to a same speaker (clustering process). Then this is the main task treated in this paper.

SD is closely related to speech processing, driven by the exponential growth of multimedia content that needs to be segmented, clustered, detected, and archived efficiently. The primary goal of SD is to segment an audio stream into homogeneous segments containing the voice of a single speaker (a process known as speaker change detection) and to group these segments by associating those belonging to the same speaker (clustering). This is the main focus of the present work.

Typically, no prior information is available regarding the number or identities of speakers in an audio stream. However, in this work, we leverage prior knowledge about the number of speakers to enhance the detection process. The speaker classification step involves grouping speaker segments within the context of audio document scenarios.

This paper presents an SD system that employs three classifiers and explores their fusion within the proposed architecture (Kenai 2019a). The detection system is examined in two configurations: one where each classifier operates independently and another utilizing a multi-classifier fusion approach.

The remainder of this paper is organized as follows:

- Section 2 presents the architecture based on Voice Activity Detection (VAD) and individual classifiers.
- Section 3 describes the proposed strategy using a multi-classifier fusion approach.
- Section 4 evaluates the performance of each method using the NIST 2005 database.
- Section 5 concludes the paper and outlines perspectives for future research.

## 2 Proposed Architecture for Speaker Detection (SD)

A Speaker Detection (SD) system consists of three main steps: 1) Acoustic Segmentation: This step detects segments containing only speech while eliminating those that contain silence, music, noise, or other non-speech elements; 2) Speaker Segmentation: This process divides the audio stream into mono-speaker segments; 3) Clustering: This step groups the segments obtained from the speaker segmentation by associating them with each target speaker.

Figure 1 illustrates the proposed SD configuration, which integrates two Voice Activity Detection (VAD) modules. The first is a temporal VAD applied before the parameterization step, and the second is a cepstral VAD applied after parameterization. This dual VAD configuration significantly enhances the performance of traditional SD systems described in the literature (Kenai 2018), (Mertens 2012), (Dupuy 2012). Further details about this configuration are available in (Kenai 2019a).

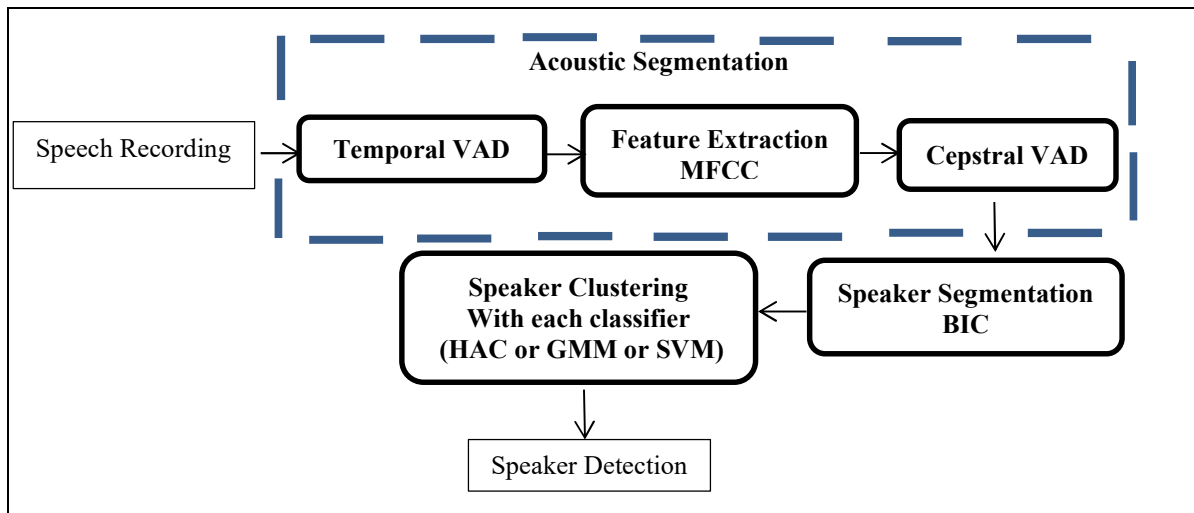


Fig.1. Proposed configuration for the SD system

## 2.1. Segmentation

The goal of audio segmentation is to get more consistent segments. Each segment has specific content: music, sound or speech. In the literature, several works on automatic speech segmentation have been proposed. The importance of this type of approach appears especially when it comes to detecting several different speakers. Other more generic segmentation techniques exploit the characteristics of audio content to segment content such as, for example, detecting audio passages containing significant silence or detection of jungles that can also infer transitions. In this work, we are interested in acoustic segmentation and speaker segmentation for the implementation of our approach (Kenai 2019a):

- **Acoustic Segmentation:** This process includes three essential phases necessary for the parameterization step: temporal Voice Activity Detection (VAD), feature extraction using Mel-Frequency Cepstral Coefficients (MFCC), and cepstral VAD.
- **Speaker Segmentation:** One common technique for speaker segmentation involves detecting speaker changes using the Bayesian Information Criterion (BIC) distance. This technique relies on the principle that two segments belonging to different speakers can be distinguished by calculating a measure of similarity between them.

## 2.2. Speaker clustering

The goal of speaker clustering in multi-speaker detection tasks is to increase the length of segments, thereby reducing the impact of shorter segment lengths on the performance of the speaker identification system (Vesnicer 2008). After completing the acoustic segmentation and speaker segmentation steps, we ideally obtain audio stream segments containing the speech of a

single speaker. The purpose of clustering is to group these segments from the previous phase according to the target speaker (Anguera 2012), (Dupuy 2012), (Delgado 2015).

In our work, the grouping process is treated as a supervised classification problem. For this purpose, a hypothesis regarding the number of speakers is assumed, and a priori information about the target speakers is available (Kenai 2018). Ideally, at the end of the clustering process, each class should exclusively contain the speech of a single speaker.

We used three classifiers in this architecture:

- **HAC Clustering:** Speaker clustering involves grouping unknown speaker utterances based on their associated speakers. The most commonly used method for speaker clustering is Hierarchical Agglomerative Clustering (HAC) (Anguera 2012), (Kenai 2018). The advantage of the HAC algorithm lies in its simplicity and ease of implementation. However, its performance heavily depends on the predefined threshold. Numerous improvements to the HAC method have been proposed in the literature. The principle of HAC is that if the utterances belonging to class A and the utterance X are the two closest among all, and the distance between them is smaller than the predefined threshold, utterance X is merged into class A.
- **GMM Classifier:** In our work, Gaussian Mixture Models (GMMs) are used following the Universal Background Model (UBM)-GMM paradigm. The estimation of the background model (UBM) parameters is performed using the Expectation-Maximization (EM) algorithm, while the speaker models (GMMs) are adapted using Maximum A Posteriori (MAP) adaptation from the UBM. Only the means are adapted, while the weights and covariance matrices remain unchanged across models. The MAP criterion allows for precise adaptation of each distribution, although it requires a relatively large amount of data to obtain a robust model estimate. This paradigm has been widely used because GMMs employ diagonal covariance matrices, simplifying computations, and MAP adaptation only modifies the mean parameters of the distributions (Kenai 2019b), (Kenai 2020).
- **SVM Classifier:** The Support Vector Machine (SVM) is a powerful discriminative classifier widely used in speaker recognition. It has been applied with spectral, prosodic, and high-level features. Currently, SVM is among the most robust classifiers in speaker verification and has been successfully combined with Gaussian Mixture Models (GMMs) to enhance accuracy. One of the key reasons for the popularity of SVM is its strong generalization ability, allowing it to effectively classify unseen data (Anguera 2012), (Kenai 2018), (Kenai 2019a), (Kinnunen 2010).

### **3 Application of Multi-Classifer Fusion for Speaker Detection (SD)**

As in other classification tasks, the combination of information from multiple sources of evidence, a technique known as fusion, has been widely applied in recognition systems (Ho

1994), (Lam 1997), (Chatzis 1999), (Verlinde 1999), (Verlinde 2000), (Bousquet 2011), (Buyssens 2011).

### **3.1. Fusion approach**

Fusion plays an important role in SD. The success of multi-classifier systems largely depends on the effectiveness of the fusion strategies applied. An appropriate fusion strategy is crucial for combining the evidence obtained from various classifiers. There are two possible stages for applying fusion strategies in multi-classifier systems: pre-matching and post-matching.

Fusion in multi-classifier systems can be accomplished at five different levels: sensor level, feature level, score level, decision level, and rank level. The first two levels (sensor and feature) are categorized as pre-matching fusion, while the last three levels (score, decision, and rank) fall under post-matching fusion.

Typically, features are first extracted from the audio stream, followed by speaker change detection. An individual classifier is applied to each feature set, and the scores or decisions are then combined. This approach means that multiple models for each speaker are stored. Additionally, fusion can also be achieved by modeling the same features using different classifier architectures. In our work, we utilized HAC, GMM, and SVM classifiers.

The fusion approach involves using the decisions from all three classifiers to match speaker segments with the target speaker in the audio stream. The decision-scoring program compares the reference segmentation with the automatic segmentation. In this case, while there is no explicit reference segmentation, classification by each classifier operates independently.

We assume that all three classifiers detect the same number of speakers in the audio stream. Fusion is then applied at the decision level, where the classifiers (using clustering with the value of  $N$  fixed a priori, as explained in Section 2.2) agree. The common segments (those on which all three classifiers or at least two classifiers agree) are retained and compared with the reference segmentation, keeping only the common segments. This process constitutes decision-level fusion.

### **3.2. Decision level fusion**

For SD, we investigated a strategy that fuses the scores at the decision level provided by each classifier (HAC, GMM, and SVM). Figure 2 illustrates the fusion architecture adopted for our system.

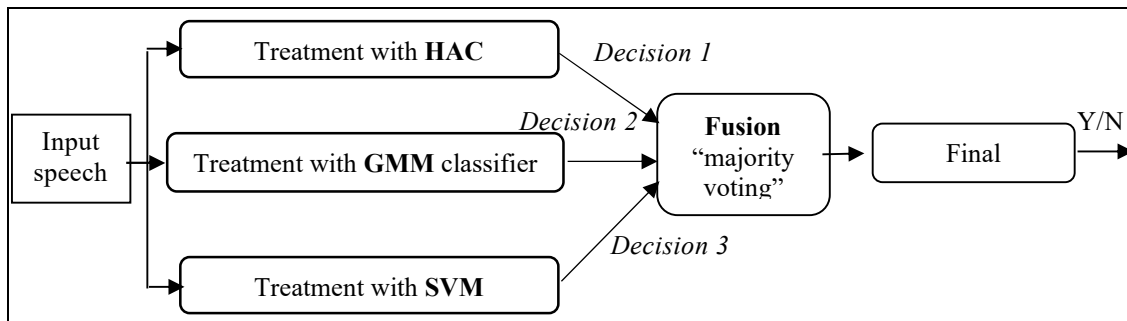


Fig.2. Parallel architecture for fusion of multi-classifiers at the decision level

- Decision 1: Obtained from a processing block that includes the following phases: temporal VAD, MFCC feature extraction, cepstral VAD, and speaker clustering using HAC.
- Decision 2: Obtained from a processing block containing the same phases but utilizing the GMM classifier.
- Decision 3: Obtained from a processing block containing the same phases but using a different classifier, the SVM.

In this fusion method, the information from multiple decision modules (classifiers) is combined to make a unified decision about the speaker's identity or detection. Each speaker is individually pre-classified, and the final classification is based on the fusion of the results from multiple classifiers.

The final output of each individual classifier can be integrated using fusion methods such as: “AND” the output is "match" only if all classifiers agree that the input sample matches the template. “OR” the output is "match" if at least one classifier determines that the input sample matches the template. “Majority Voting” the input sample is assigned to the identity on which the majority of classifiers agree. “Weighted Majority Voting”, “Bayesian Decision Fusion”, “Dempster-Shafer Theory of Evidence”, and “Behavior Knowledge Space” are other advanced fusion techniques (Verlinde 1999), (Huang 1995), (Kuncheva 2002). In our work, we selected the majority voting technique due to its effectiveness and its reliance on the consensus of multiple classifiers. Figure 3 illustrates the main configuration adopted for the SD system.

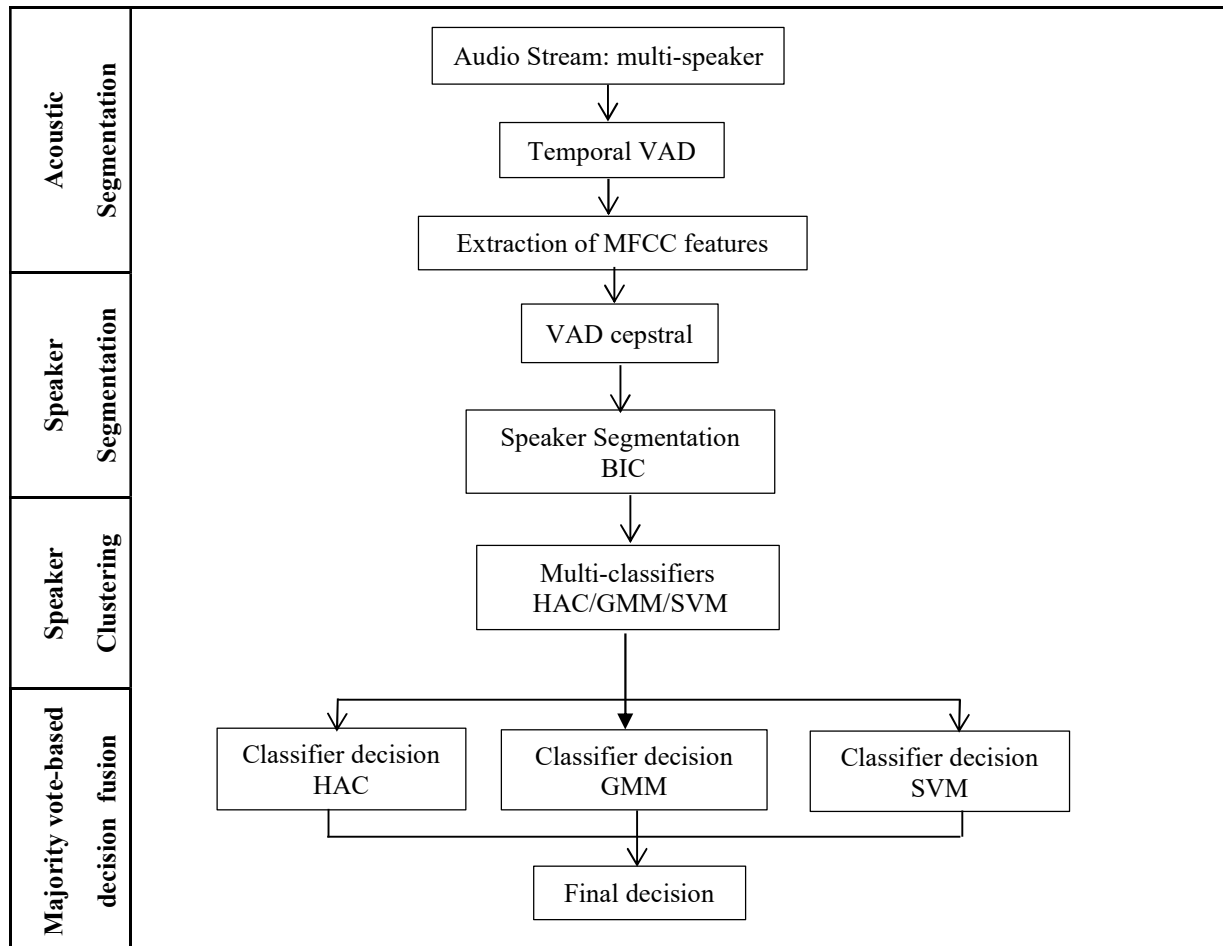


Fig.3. Proposed main configuration of an SD system using decision fusion of multi-classifiers with majority voting (Kenai 2021)

From this figure, we can distinguish 4 essential steps in the decision fusion process through voting: 1) Acoustic segmentation, which includes three phases: temporal VAD, cepstral VAD, and the extraction of acoustic features. 2) Speaker segmentation using the BIC approach. 3) Classification of the speakers. 4) Fusion of decisions using the voting technique.

In the first step, we extract features from the audio streams (scenarios) using MFCC, applying both VAD techniques to minimize silence segments (or zones). In the second step, we use the BIC technique to detect speaker changes. The third step involves classifying these speakers based on the adopted classifier type in order to construct classes or models for each speaker [6]. Finally, in the last step, we merge the decisions from the three classifiers (HAC, GMM, and SVM) to obtain a more robust final decision (Kenai 2021).

For the majority voting technique, the correct class is the one most frequently selected by the different classifiers. If all classifiers select different classes or in the event of a tie, the class with the highest overall output is chosen as the correct class. The final decision of the system is obtained through the majority vote fusion technique, as presented below:

$$Q(x) = \arg_{i=1}^K \max y_i(x) \quad (1)$$

where  $K$  is the number of classifiers and  $y_i(x)$  represents the output of the  $i^{\text{th}}$  classifier for the input vector  $x$ .

## 4 Experiments and Discussions

The experimental corpus used is an excerpt from the NIST2005 universal database, with a sampling rate of 16 kHz. The corpus is multi-speaker and multilingual, consisting of telephone recordings that were used for both training models and evaluating the speaker detection (SD) system. The setup for our experiments is as follows:

- Training phase: Six speakers were selected to form six models, with approximately 16 minutes of speech per speaker, totaling 1 hour and 35 minutes.
- Testing phase: A total of 27 test scenarios were conducted, with test durations ranging from 40 to 50 seconds, involving 1 to 5 speakers per scenario.

In this section, we present the evaluation of our SD system using two configurations, as illustrated in Figures 1 and 3. The experiments were conducted on the telephone corpus described earlier. Two series of experiments were carried out: the first using Configuration 1 and the second using our proposed Configuration 2.

### 4.1. Evaluation Results of the SD system

This section discusses the results of several speaker detection (SD) experiments. The first set of experiments focused on selecting the most appropriate classifier for our task. The performance metric used for speaker detection is the Speaker Detection Rate (SDR), which is calculated as:

$$SDR\% = \frac{\text{number of correctly detected segments of each speaker}}{\text{total number of referenced segments of each speaker}} * 100 \quad (2)$$

In scenarios where training data for each speaker in the audio stream is unavailable, the detection rate is calculated by dividing the number of correctly clustered segments (via Hierarchical Agglomerative Clustering, HAC) by the number of manually labeled reference segments. In cases where training data is available, Gaussian Mixture Models (GMMs) and Support Vector Machines (SVMs) are used instead of HAC clustering.



The results of the SDR for each classifier and speaker are presented in Figure 4. The SVM classifier outperformed both the HAC and GMM classifiers across all speakers.

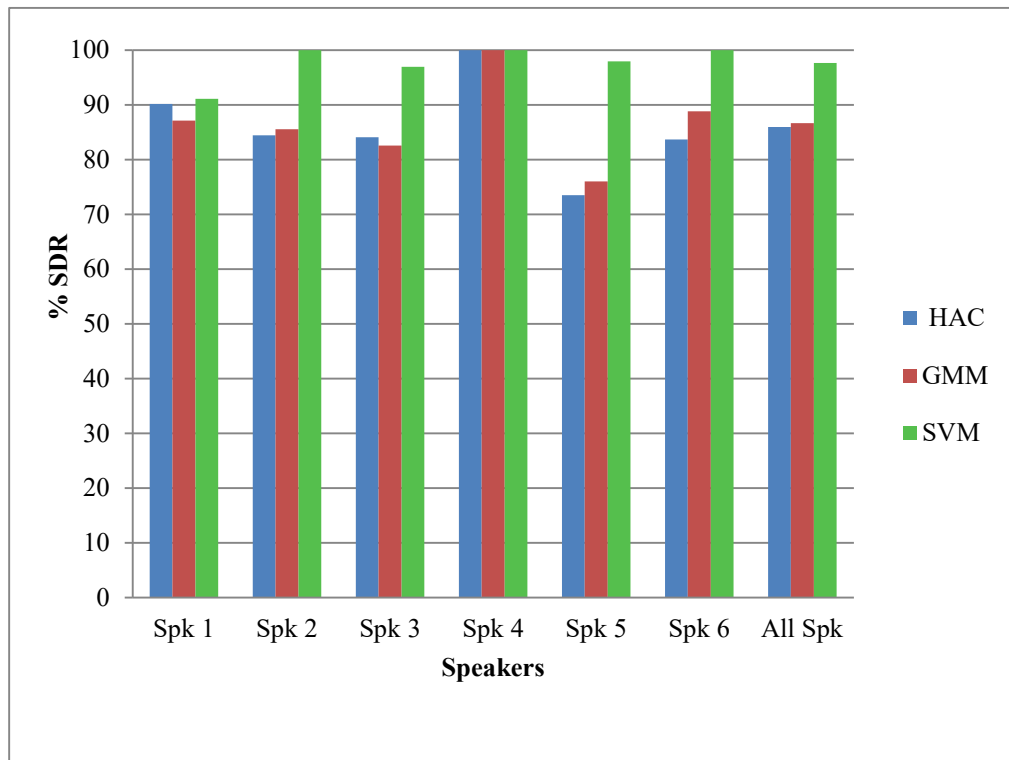


Fig.4. Performance histogram of the SDR task using HAC, GMM and SVM classifiers

To evaluate the performance of each classifier used in our SD system, we compared them as shown in Table 1.

TABLE 1. Experimental results on the SD task

Classifiers	Performances (SDR)
HAC	85.98%
GMM	86.68%
SVM	97.67%

Therefore, we interpret the performance of the individual classifiers (HAC, GMM, and SVM) as follows:

- HAC achieved an SDR of 85.98%, demonstrating its effectiveness under certain conditions. However, its performance is limited when robust training data is unavailable, as it relies solely on clustering techniques. HAC is thus better suited for scenarios where unsupervised methods are required.

- GMM, with an SDR of 86.68%, slightly outperforms HAC. This indicates that GMMs are better at modeling the acoustic variations in speech segments, particularly when training data is available. However, the improvement over HAC is modest; suggesting that more advanced methods may be needed in such contexts.
- The SVM classifier achieved an impressive SDR of 97.67%, far outperforming both HAC and GMM. This result highlights SVM's ability to maximize class separability in high-dimensional feature spaces. It is particularly effective when reliable and well-labeled training data is available, making it the best-performing individual classifier in this experiment.

The superior performance of the SVM motivated us to explore a complementary approach involving classifier fusion to further enhance the SD system's performance. The objective of this fusion approach was to maximize SDR by combining the strengths of individual classifiers.

#### 4.2. Evaluation Results of Classifier Fusion for SD

The fusion approach focuses on segments correctly detected by multiple classifiers. Figure 5 illustrates the general principle of decision-level fusion for detecting N speakers.

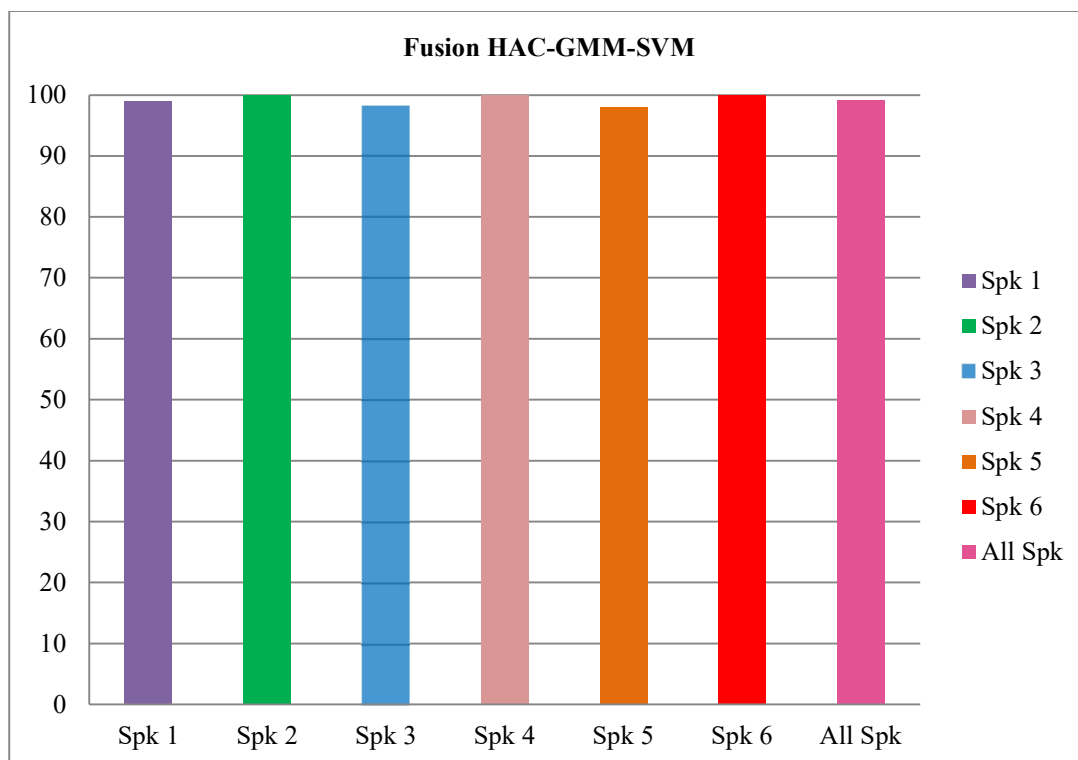


Fig.5. Performance of the system with multi-classifier fusion application

The results of the fusion approach, as shown in Table 2, highlight the improvement in SDR. The fusion of the HAC, GMM, and SVM classifiers achieved a significant performance boost, increasing the SDR to 99.18%, compared to 97.67% for the best individual classifier (SVM).

TABLE 2. Comparison of classifiers and their fusion using majority voting

Classifiers	Performances (SDR)
HAC	85.98%
GMM	86.68%
SVM	97.67%
Fusion	<b>99.18%</b>

The fusion method improves performance by combining decisions from multiple classifiers. When all classifiers agree on a segment, the decision is retained. In cases of disagreement, a majority voting strategy is employed to ensure robust decisions. This approach effectively addresses errors made by individual classifiers, achieving the highest possible SDR for the NIST2005 dataset.

The performance of this fusion approach resulted in an SDR of 99.18%, an improvement of 1.51% compared to the best individual classifier (SVM). This gain is significant in scientific contexts, where even small performance improvements can have substantial impacts on applications such as speaker recognition or detection.

Regarding the results of individual performances of the speakers, some variability is observed:

- Certain speakers, such as Speaker 4 and Speaker 6, achieved an SDR of 100% with either SVM or the fusion approach, indicating that their segments were highly distinct and easy to detect.
- Others, such as Speaker 5, showed slightly lower performance (97.98% with fusion), likely due to acoustic similarities with other speakers or noisy data.

The fusion approach improved detection across all speakers, achieving greater consistency and reducing disparities between individual speaker performances.

## 5 Conclusion

The experimental results demonstrate the effectiveness of our SD system, which integrates voice activity detection (VAD) with HAC, GMM, and SVM classifiers at the decision level. Among the individual classifiers, SVM achieved the highest speaker detection rate (SDR). However, the fusion approach clearly showed an overall improvement in performance.

These findings highlight the robustness of the proposed fusion method, suggesting its potential applicability to other speaker recognition tasks to further enhance performance. Moreover, this fusion-based method could be extended to other speaker recognition systems, improving both reliability and accuracy.

## References

- (Anguera 2012) Anguera X., Bozonnet S., Evans N., Fredouille, C., et al. "Speaker diarization: A review of recent research". IEEE Transactions on Audio, Speech, and Language Processing, 20 (2), 356-370, 2012.
- (Bousquet 2011) Bousquet, P. M., Matrouf, D. & Bonastre, J. F. "Intersession compensation and scoring methods in the i-vectors space for speaker recognition". In Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH). ISCA, 2011.
- (Buysens 2011) Buysens P. "Fusion de différents modes de capture pour la reconnaissance du visage appliquée aux e\_transactions". Thèse de doctorat, Université de Caen Basse Normandie, 2011.
- (Chatzis 1999) Chatzis V., Bors A., & Pitas I. "Multimodal decision-level fusion for person authentication". IEEE Transactions on Systems, Man and Cybernetics, Part A : Systems and Humans, Vol. 29, No. 6, pp. 674–681, November 1999.
- (Delgado 2015) Delgado H., Anguera Miro X., Fredouille C., & Serrano J. "Fast singleand cross-show speaker diarization using binary key speaker modeling". IEEE Transaction on Audio, Speech and Language Processing, vol. 23, pp. 2286–2297, 2015.
- (Dupuy 2012) Dupuy G., Rouvier M., Meignier S., & Estève Y. "Segmentation et Regroupement en Locuteurs d'une collection de documents audio". Cross-show speaker diarization. In Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 1: JEP, pp. 433-440, June 2012.
- (Ho 1994) Ho T. K., Hull J. J., & Srihari S. N. "Decision combination in multiple classifier systems". IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 16, N° 1, pp. 66–75, 1994.
- (Huang 1995) Huang Y.S. & Suen C.Y. "A method of combining multiple experts for the recognition of unconstrained handwritten numerals". IEEE transactions on pattern analysis and machine intelligence, vol. 17, N° 1, p. 90-94, 1995.
- (Kenai 2018) Kenai O., Asbai N., Ouamour S., Guerti M., & Djeghiour S. "Speaker diarization and detection system using a priori speaker information". In 2nd International Conference on Natural Language and Speech Processing, pp. 1-6, IEEE, 2018.
- (Kenai 2019a) Kenai O., Ouamour S., Guerti M., Asbai N. "A new architecture based VAD for speaker diarization/detection systems". International Journal of Speech Technology, vol. 22, N° 3, pp. 827-840, 2019.
- (Kenai 2019b) Kenai, O., Djeghiour, S., Asbai, N., & Guerti, M. "Forensic gender speaker recognition under clean and noisy environments". Procedia Computer Science, 151, 897-902, 2019.
- (Kenai 2020) Kenai, O., Ouamour, S., & Guerti, M. "Impact of a Voice Trace for the Detection of Suspect in a Multi-Speakers Stream". p.199-208, ICICNIS 2020.
- (Kenai 2021) Kenai, O. "Application de la fusion multi-classifieur à la détection de locuteurs dans les conversations audio", PhD thesis, USTHB 2021.
- (Kinnunen 2010) Kinnunen T. & Li H. "An overview of text-independent speaker recognition: from features to supervectors". Speech communication, 52 (1), 12-40, 2010.
- (Kuncheva 2002) Kuncheva L. I. "A theoretical study on six classifier fusion strategies". Transactions on Pattern Analysis & Machine Intelligence, 2002, Vol. 4, N° 2, pp. 281-286.
- (Lam 1997) Lam L. & Suen C. Y. "Application of majority voting to pattern recognition : An analysis of its behavior and performance". IEEE Transactions on Systems, Man, and Cybernetics, 27, 1997.
- (Mertens 2012) Mertens R., Huang P.S., Gottlieb L., Friedland G. et al. "On the applicability of speaker diarization to audio indexing of non-speech and mixed non-speech/speech video soundtracks". International Journal

- of Multimedia Data Engineering and Management (IJMDEM), 3 (3), 1-19, 2012.
- (Reynolds 2004) Reynolds, D. A., & Torres-Carrasquillo, P. "The MIT Lincoln Laboratory RT-04F diarization systems: Applications to broadcast audio and telephone conversations". In Proc. Fall 2004 Rich Transcription Workshop (RT-04), 2004.
- (Shon 2019) Shon S., Dehak N., Reynolds D., & Glass J. "Mce 2018: The 1st multi-target speaker detection and identification challenge evaluation". arXiv preprint arXiv:1904.04240, 2019.
- (Verlinde 1999) Verlinde P., Acheroy M. et al. "A contribution to multi-modal identity verification using decision fusion". In : Proc. PROMOPTICA. pp. 1-16, 1999.
- (Verlinde 2000) Verlinde, P., Chollet, G., & Acheroy, M. "Multi-modal identity verification using expert fusion". Information fusion, 1 (1), 17-33, 2000.
- (Vesnicer 2008) Vesnicer B., Mihelic F., & Zibert J. "Development of a speaker diarization system for speaker tracking in audio broadcast news: a case study". Journal of Computing and Information Technology, 16(3), 183-195, 2008.