Enhancing Speaker Recognition with Multiple Observations over Mobile Networks

Igor Bisio, Chiara Garibotto, Fabio Lavagetto and Andrea Sciarrone
University of Genoa - DITEN Department
Via all’Opera Pia 13
16145, Genova Italy
Emails: igor.bisio@unige.it, chiara.garibotto@edu.unige.it, fabio.lavagetto@unige.it, andrea.sciarrone@unige.it

Abstract—Nowadays, the widespread use of Mobile Devices (MDs) opens the door to exploit the presence of multiple nodes to accomplish collaborative tasks. In this paper, a speaker recognition system for MDs based on a multiple-observations approach is presented. We propose different fusion and clustering algorithms aimed at efficiently exploiting signals coming from multiple sensors. Numerical results show that in most cases our multiple-observations approach is able to significantly improve the performance of a single-receiver approach.

I. INTRODUCTION

Nowadays, Mobile Devices (MDs) allow different types of processing. For example, modern smartphones, equipped with powerful microphones for the purpose of speech processing, continuously offer new speech-audio functions for the emerging context-aware services in Mobile Wireless Networks (MWNs). Thanks to their great capabilities and their great diffusion, one of the most interesting problem in MWNs is to find an efficient way that allows MDs to cooperate together for a common goal. On the other hand, understanding audio context (e.g., finding out the number of speakers rather than recognizing who is speaking) is one of the most interesting problems in the related literature which has many practical applications ranging from security to advertisement.

The aim of this paper is to enhance the accuracy (i.e., the percentage of correct recognition) of a system able to recognize the identity of a speaker, among a group of known speakers, thanks to a signal processing algorithm that exploits multiple observations of the speech signal acquired by different smartphones. In many practical cases a speaker talks in an environment in which many MDs are present, for example during a presentation in conference or during a lecture in a classroom. For this reason, enhancing the accuracy of a speaker recognition system by exploiting multiple observations from different MDs may represent an important tool that can be extremely useful in several realistic scenarios.

When multiple observations of an audio signal coming from different devices are available, the need of efficiently managing the gathered information arises. In the literature, many works describe the concept of multiple observations and data fusion. Some of them deal with the problem of merging data coming from heterogeneous sensors, such as cameras, microphones, and accelerometers [1], [2]. Other works instead apply the concept of fusing observations from different but homogeneous sensors. Many literature works consider the multiple observations problem applied to different types of sensors [3] and [4]. An interesting approach is described in [5], where a novel concept of data fusion named service-controlled networking is proposed.

Another approach that can be used when dealing with multiple sensors is node clustering. The main criteria on which to base the subset selection strongly depend on the considered application. In particular, [6] defines a strategy that allows to select sensors based on the quality of the resulting data. Also [7] describes a selection scheme that proposes a trade off between reliability of the system and power consumption. Other works use multiple observations for classification applied to tracking and event detection, as presented in [8].

In this paper, we propose a set of methods to efficiently use the available information coming from different observations of the speech signal acquired with multiple MDs. We analyse the performance of a speaker recognition approach based on multiple observations over MDs with respect to an approach that only employs a single observation scheme. The key aspect of this work is the definition of different fusion and clustering criteria that allow conveniently exploiting the information coming from multiple MDs. The rest of this paper is organized as follows: Section II describes the speaker recognition algorithm, the proposed multiple observation-based methods are described in Section III, whereas the obtained numerical results are presented in Section IV. Finally, conclusions are drawn.

II. SPEAKER RECOGNITION ALGORITHM

In the literature, many works address the problem of speaker recognition. Among others, valuable ideas are presented in [9], [10]. In order to train systems and identify the correct speaker, a wide range of speech signal features has been proposed: Mel-Frequency Cepstral Coefficients (MFCC) [11], Perceptual Linear Prediction (PLP) and Linear Predictive Cepstral Coefficients (LPCC) [12], [13]. Another important aspect in speaker recognition systems lies in the classification tool to employ for the recognition task. Gaussian Mixture Model (GMM) [14] and Support Vector Machine (SVM) [11] are the most used. In many works, Hidden Markov Models
(HMM) and the Artificial Neural Networks (ANN) [15] are also commonly employed.

In this work, the audio signal is filtered with a second-order Butterworth BPF with a bandwidth \( B \in [50, 500] \) [Hz]. We divide the signal into short segments called frames, during which the speech signal can be considered as stationary. In our case, each frame has a length of \( T = 40 \) [ms] and it is selected so that two consecutive frames are overlapped for one third of their duration. After the framing block, the features are computed. Exploiting the Weka environment [16] we have performed feature selection to understand the best feature set to employ. We started from 13 MFCC, Delta, DeltaDelta and the respective 91 Shifted-Delta and Shifted-DeltaDelta for a complessive number of 221 features. The selection on the best set of feature has been performed by applying the Information Gain criterion. This results in the selection of a set of 48 features, composed of a mix of MFCC and Shifted-Delta coefficients. Once the features have been extracted, they will be employed, together with an SVM classifier, in two separate phases: i) training phase and ii) decision (testing) phase.

A. Training phase

We define \( \mathbf{X}_f \) as the feature vector for the \( f \)-th frame. \( \mathbf{Y} \) is the vector containing all the classes (i.e., the names of all speakers). Let \( |\mathbf{Y}| \) be the total number of considered speakers and \( h \in [1, |\mathbf{Y}|] \) be the classes’ index. \( \check{Y}_h \) is the \( h \)-th class (i.e., the single speaker’s name). Given the quantity \( \mathbf{Y}^f \in \mathbf{Y} \) which is the class corresponding to the vector \( \mathbf{X}_f \), the association \( (\mathbf{X}_f, \mathbf{Y}^f) \), \( \forall f \in [1, F] \) is called observation.

The SVM is trained by using the One-Against-One (OAO) method. It relies on constructing classifiers where each one is trained by using data from two classes. It trains different hyperplanes by using combinations of all the \( |\mathbf{Y}| \) classes taken in pairs. This leads to obtain \( S_{\text{OAO}} = \{ (\mathbf{Y}^f) = \mathbf{Y} \mid i \neq j \} \) different hyperplanes. The single SVM, built for the classes \( Y_h \) and \( Y_k \), can be obtained by computing the aforementioned hyperplane that can be expressed as a function of its orthogonal vector. Further detail about the analytical description of the SVM have been omitted for the sake of simplicity.

The SVM performs classification based on a so called Kernel function. In the literature, several kernel functions have been proposed. One of the most common is the Radial Basis Function (RBF) or Gaussian Kernel which is the same non-linear kernel function employed in this paper. The SVM training parameters, defined in [17], \( C \) (i.e., the complexity constant) and \( \gamma \) (i.e., the kernel parameter) have been empirically set to \( C = 0.5 \) and \( \gamma = 0.01 \), respectively.

We denote with \( F_h \) and \( F_k \) the total number of frames associated to the classes \( Y_h \) and \( Y_k \) contained within an audio file, respectively. The scalars \( \check{Y}_f \in [-1, 1] \) are the numeric binary labels associated to the semantic labels \( Y_f \) of the feature vectors \( \mathbf{X}_f \). They are defined as:

\[
\check{Y}_f = \begin{cases} 
1, & \text{if } Y_f \equiv Y_h \\
-1, & \text{if } Y_f \equiv Y_k 
\end{cases} \tag{1}
\]

B. Decision (Testing) phase

Defining with \( |\mathbf{X}_f| \) the number of features for each frame and with \( F \) the total number of frames contained within an audio file, the \( F \times |\mathbf{X}_f| \) matrix \( \mathbf{\omega} \), containing all the feature vectors for an audio file is:

\[
\mathbf{\omega} = [\mathbf{X}_1 \ldots \mathbf{X}_f \ldots \mathbf{X}_F]^T. \tag{2}
\]

Every time that a speaker must be recognized, the SVM outputs a \( F \times 2 \) matrix \( \mathbf{\Omega}_{Y_h,Y_k} \{ \mathbf{\omega} \} \), called Probability Matrix (PM), one for each trained hyperplanes. The number of PMs is the number of all the possible combinations (i.e., the order doesn’t matter) of all the considered speakers taken in pairs. Each element \( \mathbf{\Omega}_{Y_f,Y_k} \), \( f \in [1, F], t \in \{1, 2\} \) of this matrix is the \( a\text{-posteriori} \) probability of the \( f \)-th feature vector \( \mathbf{X}_f \) of belonging to the class identified by the binary label \( \check{Y}_f \), as reported in Eq.3.

\[
\mathbf{\Omega}_{Y_h,Y_k} \mathbf{\{ \omega \}} = \begin{bmatrix} 
Pr \left( \check{Y}_f = 1 | X_f \right) & Pr \left( \check{Y}_f = -1 | X_f \right) \\
\vdots & \vdots \\
Pr \left( \check{Y}_f = 1 | X_F \right) & Pr \left( \check{Y}_f = -1 | X_F \right) 
\end{bmatrix} \tag{3}
\]

The index \( t \in \{1, 2\} \) is obtained from the semantic labels \( Y_f \) as follows:

\[
t = \Phi \left( Y_f, \mathbf{\Omega}_{Y_h,Y_k} \right) ^2 - \left[ \Phi \left( Y_f, \mathbf{\Omega}_{Y_h,Y_k} \right) / 2 \right] + 1, \tag{4}
\]

where the expression \( \Phi \left( Y_f, \mathbf{\Omega}_{Y_h,Y_k} \right) \) is a function that takes as inputs the semantic label \( Y_f \) and the PM \( \mathbf{\Omega}_{Y_h,Y_k} \). It returns the binary label \( \check{Y}_f \in \{-1, 1\} \) associated to the semantic label \( Y_f \). For example, \( \Phi \left( Spk_A, \mathbf{\Omega}^{Spk_A,Spk_B} \right) = 1 \) while \( \Phi \left( Spk_B, \mathbf{\Omega}^{Spk_A,Spk_B} \right) = -1 \). The single frame \( f \) is associated to the predicted class \( Y_k \in \mathbf{Y} \) by computing the index \( k^* \) as follows:

\[
k^* = \arg \max_k \left[ \frac{1}{(|\mathbf{Y}| - 1)} \sum_{h=1}^{|\mathbf{Y}|} \mathbf{\Omega}_{Y_f,Y_h} \right] \forall k \in [1, |\mathbf{Y}|]. \tag{5}
\]

Iterating Eq. 5 for all the \( F \) frames present in an audio file, allows building an \( 1 \times F \) vector \( \mathbf{L} \) which contains all the predicted classes. The single \( L_f \) element of the vector \( \mathbf{L} \), \( f \in [1, F] \), is the speaker’s label associated to the single frame \( f \). Finally, from the vector \( \mathbf{L} \), an \( 1 \times |\mathbf{Y}| \) scoring vector \( \mathbf{S} \) is built. It contains all the scores \( s_h, h \in [1, |\mathbf{Y}|] \) of the speakers involved in the recognition process. Every value \( s_h \) represents the likelihood of the considered audio file of belonging to the speaker \( Y_h \). It is computed by using the pseudo-code reported in Algorithm 1. The final decision related to who is speaking is simply taken with the Best-Win approach: the recognized speaker is the one who has the highest score.
III. MULTIPLE-OBSERVATIONS APPROACH

To improve the speaker classification accuracy through a collaborative approach, we decided to exploit more than one MD receiver. We propose different algorithms aimed at opportunistically managing the contributions of the available MDs, based on the conditions of each receiver. We develop 4 different ways to manage multiple-observations, that aim at obtaining an overall result by jointly exploiting the signals captured by the receivers.

A. Arithmetic Mean Fusion (avg)

The first scheme performs the arithmetic mean of the scores obtained at the different MDs. Each receiver performs the classification independently of the other MDs and it produces an output vector $S$, whose elements refer to the scores of the speakers belonging to the speaker set. The final score vector $E[S]$, obtained by applying the fusion criterion, is the following:

$$E[S] = \frac{1}{N} \sum_{i=1}^{N} S_i,$$

where $N$ is the total number of receivers and $S_i$ is the score vector produced by the $i$-th receiver.

B. Weighted Average Fusion (snr2OverDist)

The second approach computes a weighted average of the contributions of the various receivers, by taking into account attenuation and noise conditions related to the specific MD. As confirmed by the single observation analysis reported in Section IV-B, the classification accuracy decreases as the source-to-destination distance increases and as the signal-to-noise ratio decreases. We define a parameter $p_i$ that is proportional to the SNR and to the distance, such that:

$$p_i = \frac{(snr_i)^\xi}{d_i},$$

where $snr_i$ is the signal-to-noise ratio affecting the $i$-th receiver and $d_i$ is the distance between the source of the audio signal and the $i$-th MD. The quantity $\xi$ is a power factor that can enhance the contribution of the noise with respect to the attenuation. The actual weighting factor $w_i$ is obtained by normalizing the parameters computed for the different nodes:

$$w_i = \frac{p_i}{\sum_{i=1}^{N} p_i}.$$

The overall output scoring vector $E_{w}[S]$ resulting from this approach is therefore:

$$E_{w}[S] = \sum_{i=1}^{N} w_i \cdot S_i,$$

which is the weighted sum of the contributions obtained by each receiver.

C. Parameter-based Clustering (snr2OverDist-clusterP)

As the MDs’ conditions become more challenging, the contribution of the observations will be more noisy, so affecting the overall classification performance. To efficiently tackle this issue, we also propose two clustering methods that aim at exploiting only a subset of all the available receivers. Considering the parameters defined earlier, we select the MDs that will contribute in recognizing the speaker based on predefined thresholds.

The first clustering algorithm computes the parameter $p_i$ for each available MD as reported in Eq. 7 and then compares this value to a threshold $\epsilon_p$. If it is lower than the threshold, the $i$-th node is turned off and its contribution is not considered for computing the final scoring vector. In formula:

$$w_i = \begin{cases} \frac{p_i}{\sum_{i=1}^{N} p_i}, & \text{if } p_i \geq \epsilon_p \\ 0, & \text{otherwise.} \end{cases}$$

D. Weight-based Clustering (snr2OverDist-clusterW)

The last clustering algorithm proposes an alternative selection criterion, which consists in using the weighting factor $w_i$ as clustering parameter: if $w_i$ is lower than a pre-defined threshold $\epsilon_w$, the node is excluded from the computation of the final score.

$$w_i = \begin{cases} \frac{p_i}{\sum_{i=1}^{N} p_i}, & \text{if } w_i \geq \epsilon_w \\ 0, & \text{otherwise.} \end{cases}$$

IV. NUMERICAL RESULTS

A. Tests conditions

MDs are simulated and implemented within the MatLab computing environment on a personal computer running Linux OS. The database used for all tests consists in a set of clean audio signals belonging to 4 different speakers producing several utterances in English language. The position of the audio source signal (i.e., the speaker) is known a-priori. This is not
The average accuracy of the system can be defined as:

\[
\Theta(m, f) = \begin{cases} 
1, & \text{if the } f\text{-th frame of the } m\text{-th file has been correctly classified;} \\
0, & \text{otherwise.}
\end{cases}
\]

The average accuracy of the system can be defined as:

\[
\alpha = \frac{1}{M} \sum_{m=1}^{M} \frac{1}{F} \sum_{f=1}^{F} \Theta(m, f).
\]

This quantity has been chosen as performance parameter to compare the classification results of the proposed algorithms.

### B. Single Observation

The first analysis shown in this paper is related to the performances of the system when only a single observation is considered. This represents a starting point that can be used to evaluate the efficiency of the proposed multiple observations approaches. We ran several performance tests in order to understand the effects of attenuation and noise level on the speaker classification accuracy. In this first scenario, only one MD is involved. Figure 2 shows the results expressed as a percentage of the classification accuracy (\( \alpha \)) versus a progressive decreasing of the signal-to-noise ratio \( snr \in \{50, 45, 40, 35, 30\} \) [dB], at different source-receiver distances \( d \in \{1, 4, 7, 10\} \) [m]. The performances of the system decrease when the distance between source and destination increases.

Furthermore, the system performances strongly decrease also when the SNR decreases: if it drops under 30 [dB] the system performances become unsatisfactory (with an average accuracy of about 55%), independently of the considered distance, and the system outcome is unpredictable as the recognition is strongly corrupted by noise.

### C. Multiple Observations

We ran several tests in order to understand the behaviour of the proposed approaches in an exhaustive range of conditions. All simulations are performed with the same characteristics as described above. In all performance tests we consider up to \( N = 5 \) receivers lying at different distances from the audio source \( d \in \{1, 4, 7, 10\} \) [m] and experiencing a signal-to-noise ratio \( snr \in \{50, 45, 40, 35, 30\} \) [dB]. Since the proposed algorithms exhibit very similar behaviours with respect to the number of MDs for all the tested scenarios, for the sake of brevity we only report the results related to the case when 3 receivers are involved.

An audio signal produced by a speaker belonging to the speaker set is received and classified frame by frame at

![Fig. 1: The multiple observations approach.](image-url)
Fig. 3: Accuracy matrices obtained with the different algorithms (3 receivers scenario, $snr_3 = 50$ [dB] and $d_3 = 1$ [m])

each receiving MD independently, so that the respective score vectors $S_i$ are produced. The final score is then obtained by applying one of the proposed fusion or clustering algorithms. Consequently, the classification accuracy for that specific method is computed as stated in Eq. 14. Figure 3 shows a set of shaded matrices representing the accuracy obtained by using the different fusion and clustering algorithms when 3 MDs are involved. To better understand the obtained performance, we show the results when the conditions of the first and second node vary whereas the parameters of the third node are fixed ($snr_3 = 50$ [dB] and $d_3 = 1$). The pictures exhibit the classification accuracy as a shade of gray: the lighter the shade, the higher the percentage of correct recognition, as stated in the provided color bar.

The first tests were performed by using the avg fusion method, which consists in considering the contribution of each receiver by computing the average mean of the scores produced by each receiver (Eq. 6). Its average accuracy is depicted in Figure 3a. This approach works fine when both receivers exhibit good conditions in term of signal-to-noise ratio and attenuation, but it fails if the conditions get more challenging.

In order to avoid this behaviour, we apply the snr2OverDist fusion method. It takes into account a-priori knowledge about noise and distance in order to appropriately weight the contributions of each receiver. This method is based on using the parameter defined in Eq. 7, with $\xi = 2$. The results obtained using this algorithm are presented in Figure 3b. The image shows that this method is able to improve the classification accuracy in most cases, and in particular the performance enhancement is more evident when the conditions of the receivers are more challenging.

If all the receivers are strongly affected by noise and the signal experiences huge attenuation, the classification accuracy decreases critically. For this reason we propose, as an alternative to the weighting solution, a cluster selection of the nodes based on the previously described a-priori information.

The first clustering scheme uses the selection criterion defined in Eq. 10, with $\xi = 2$, and is called snr2OverDist-clusterP. The main difference between this approach and the
weighted average is that, when a node is under threshold, its contribution is ignored. Figure 3c shows the classification accuracy obtained by applying this criterion, with an \( a\)-priori clustering threshold \( \epsilon_{w} \) empirically set to 200. From these results it is possible to infer that this clustering algorithm is able to perform better, on average, than both the previously proposed approaches. The second proposed clustering method uses the weighting factor \( w \) as clustering parameter and for this reason it is called snr2OverDist-clusterW. This approach is defined in Eq. 11 and is computed by setting \( \xi = 2 \) in Eq. 7. The clustering threshold \( \epsilon_{w} \) used to exclude the nodes in worse conditions is empirically set to 0.3. The accuracy performances obtained by applying this method are shown in Figure 3d.

It is evident that this clustering algorithm exhibits a quite different behaviour with respect to the previously presented approaches. This is explained by the fact that in this approach the contribution of the MDs is computed with respect to the other available receivers. Therefore, when the conditions of a receiver are too challenging with respect to the other available MDs, the respective observation is excluded from computation the final score vector. This allows obtaining good results when the scenario includes an MD which lies in very challenging conditions, as this node will be identified and turned off.

**CONCLUSIONS**

In this paper a multiple-observation approach for audio speaker recognition on mobile devices has been proposed. We presented different fusion and clustering algorithms in order to exploit signals coming from multiple receivers so enhancing the speaker classification performance with respect to a single device case.

The numerical comparison of the accuracy obtained by applying the different methods in the case of 3 receivers confirms that almost in all cases the multiple-receiver approaches can achieve higher classification accuracy values with respect to those obtained when a single device is employed.

**REFERENCES**