Circuits, Systems, and Signal Processing (2019) 38:3743–3761
https://doi.org/10.1007/s00034-019-01026-z

Speaker Identification for OFDM-Based Aeronautical Communication System

Sara Sekkate[1] · **Mohammed Khalil**[1] · **Abdellah Adib**[1]

Received: 19 October 2017 / Revised: 3 January 2019 / Accepted: 4 January 2019 /
Published online: 14 January 2019

Abstract

Although a lot of research has been done on speaker identification in the presence of noise and channel variation, to the best of our knowledge, no work has been reported for aeronautical applications. In this paper, we aim to fulfill this goal by developing a Speaker Identification System (SIS) for future aeronautical communications systems. Furthermore, we present a novel feature extraction scheme based on multi-resolution analysis. The proposed features called SMFCC use Mel Frequency Cepstral Coefficients (MFCCs) features of stationary wavelet transform sub-bands. The extracted features are modeled using the i-vector approach, and support-vector machines are adopted as a back-end classifier. The performance of the proposed SIS is evaluated using two publicly available databases. Comparison of the proposed approach with the baseline MFCC feature extraction shows the feasibility and the robustness of the proposed method. Besides the noise reduction, the identification accuracy is improved by about 12% at higher signal-to-noise ratios and reaches 97.33% as compared to 88.33% using MFCC for ATCOSIM database.

Keywords Speaker identification · MFCC · SWT · Air Traffic Control (ATC) · i-vector

1 Introduction

Speaker recognition is a well-established research problem and has found use in many applications, including voice authenticated bank transactions which is referred to as

✉ Sara Sekkate
sarasekkate@gmail.com

Mohammed Khalil
medkhalil87@gmail.com

Abdellah Adib
adib@fstm.ac.ma

[1] Team Networks, Telecoms and Multimedia, LIM@II-FSTM, B.P. 146, 20650 Mohammedia, Morocco

telephone banking, access control, prison call monitoring and voice mail [13,30,31,38]. However, research has been rarely devoted to the integration of automatic speaker recognition (ASR) in the aeronautical industry.

In Air Traffic Control (ATC), voice communication serves as the main media for delivering instructions and important information between pilots and controllers. The road to improving safety in ATC, therefore, definitely passes through improving the air–ground communications safety. Air–ground communication is defined as a two-way communication between aircraft and ground stations. By far, the most prominent issue surrounding these communications is the heightened risk of callsign confusion. Callsign confusion occurs when aircraft operate in the same ATC sector with similar callsigns. As a consequence, it is possible for the pilot to accept clearances meant for others, leading to wrong subsequent actions and incidents with a high potential to cause death [23]. In this context, there are few works that can help further mitigate callsign confusion. In [21,24,42], authors have proposed the use of watermarking techniques, which consists of embedding digital aircraft identification data within the speech signal sent by the speaker before it is transmitted over the very high frequency or L-band communication channels. Hence, once the communication is established, the embedded data are automatically displayed at the receiver side allowing an automatic identification of the talker.

Automatic speaker recognition is performed in both of verification and identification modes. In verification mode, a reject/accept decision is made for each input speech to verify whether it corresponds to a claimed identity. While in identification mode, the goal is to identify an input speech by selecting one model from the previously enrolled speaker models. In [34,49], the authors have proposed a speaker verification system which combines the use of an Aircraft Identification Tag (AIT) and speaker segmentation techniques. Verification is performed by comparing some feature characteristics extracted from pilots' voices to the ones enrolled for the same AIT signature. This solution could effectively verify whether the speaker had changed or not; however, no information about speakers is provided. In [43], the integration of speaker identification system had been studied in the aeronautical context, especially in L-band Digital Aeronautical Communication System 1 (LDACS1) which is an OFDM-based aeronautical communication system. A SIS has been implemented using aeronautical noise-corrupted speech signals. Four conventional features were used including Linear Predictive Cepstral Coefficients (LPCC), Perceptual Linear Prediction (PLP), Gammatone Frequency Cepstral Coefficients (GFCC) and Mel Frequency Cepstral Coefficient (MFCC). For classification purposes, Support-Vector Machines (SVM) were considered.

The development of Speaker Recognition Systems (SRS) for real-world applications is a challenging task. Depending on the intended application, several robustness issues are available which make SRS vulnerable [56]; the two most common issues are noise and channel variations. To address individual challenges, various solutions have rolled out. Most of the existing methods to approach noise contamination focus on employing speech enhancement and noise removal techniques. In [1], spectral subtraction has been combined with empirical mode decomposition to improve the quality of AWGN-corrupted speech signals prior to speaker identification. In [19], ten different types of noise from the NOISEX-92 database [5] have been used to test the efficacy of adaptive Bionic wavelet shrinkage for noise suppression. In [7], Wiener filtering has been added as an additional step in MFCC computation process for handling the noise in speech signals from the NOIZEUS corpus [26]. In [45], relative spectra have been added as a preprocessing step prior to feature extraction, where MFCC features have been extracted from noisy speech signals recorded in a classroom environment. In [54] and [55], computational auditory scene analysis has been applied to deal with noisy speech, where signals from the 2002 NIST corpus [35] have been mixed with multi-talker babble noise, speech shape noise and factory noise. On the other hand, channel mismatch has also been dealt by, for example, using Cepstral Mean Subtraction [2] or training speaker models in various noisy conditions to reduce the mismatch between training and testing data [32]. Alternatively, new paradigms such as i-vector [12] and deep neural networks-based system named x-vector [48,53] have been introduced for compensating the variability caused by different channels and sessions.

Another approach to tackle noise robustness issues consists of proposing more effective features. In speaker recognition, the most extensively used features are MFCCs due to their ability to characterize a large amount of data with few features as well as their satisfactory performance in clean environments. However, they are sensitive to noise. In this context, some notable features were introduced such as Mean Hilbert Envelope Coefficients [40] and GFCC [47]. In [52], features were extracted from steady vowel segments. Due to the high-signal energy in those regions, speaker-specific information may be less affected by noise. Alternatively, in [27,46], MFCC features were extracted from transform domain rather than time one using Discrete Cosine Transform (DCT), Discrete Sine Transform (DST) and Discrete Wavelet Transform (DWT). It is shown that the usage of such transforms prior to the feature extraction process leads to a performance enhancement in SIS.

In order to deal with real-world problems, we present in this paper a novel SIS for the en route airspace in the future aeronautical communication system. It extends the previous research in [43], where conventional features were tested in an aeronautical environment, to incorporate more robust features. Motivated by the success of Stationary wavelet Transform (SWT) for noise reduction in [36], a novel multi-resolution feature extraction process is proposed based on conventional MFCC features and wavelet analysis. Moreover, in order to approach real-life conditions, the transmission channel as well as the background noise effects is included. Further, a comparison study with baseline MFCC features will be then conducted.

The outline of the paper is structured as follows. The next section presents an overview of a speaker identification system. Section 3 details the proposed approach. The experiments and obtained results are presented in Sect. 4. At the end of this paper, a conclusion is presented.

2 Overview of Automatic Speaker Identification System

Speaker recognition can be classified into speaker verification and speaker identification. Speaker verification allows to decide whether an unknown speaker is the one he claims to be, and speaker identification aims to identify an unknown speech sample by selecting one model from a set of enrolled speaker models. Both systems can be

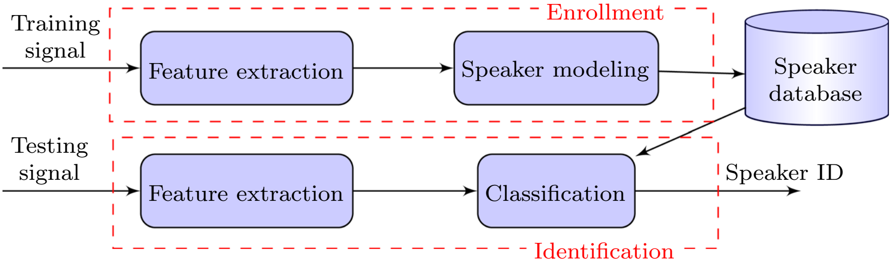

Fig. 1 General architecture of a closed-set speaker identification system

categorized as text dependent or text independent [18,22]. In text-dependent mode, the speaker provides the same utterances for both training and testing trials. In case of text-independent mode, no restrictions are imposed on spoken phrases, and hence the speaker can utter any word in order to be recognized. Both systems can be further subdivided into closed set and open set [4]. Closed-set systems suppose that the unknown speaker to be identified is known a priori to be one of the registered speakers set. In this case, when it comes to identifying an unknown speech signal, the test input speech is compared with all the available speakers in the database and the identity of the model with the closest match is returned. On the other hand, open-set systems include the possibility that the unknown speaker is none of the registered speakers. In this case, the speaker is considered as an impostor or an outsider if there is no match.

In this paper, we aim to develop a closed-set text-independent SIS that could be used to prevent callsign confusion and hence increase flight safety. The general structure of a closed-set SIS is illustrated in Fig. 1.

The identification process involves two main stages: enrollment and identification. During the enrollment stage, feature extraction is performed after the acquisition of training speech signals; it aims at providing a useful representation of the input signals in such a way that it can be understandable by the system. Based on these features, speaker models are built and stored in a database to be used in the next stage. At the identification level, classification is performed with the objective of finding the identity of the test speaker by comparing the extracted features with all the speaker models stored in the training stage.

2.1 Speaker Modeling

Speaker modeling is intended to find a representation of the speaker's characteristics in such a way that he can be distinguishable from all other speakers.

The i-vector modeling approach was developed in [12] and has risen to prominence as a state of the art in speaker recognition. The i-vector aims at modeling both speaker and channel variability by generating a low-dimensional space named total variability space. The i-vector extraction procedure can be depicted in Fig. 2.

There are three types of speech data used for the i-vector-based model. The background or development data contain a large amount of speech spoken by different

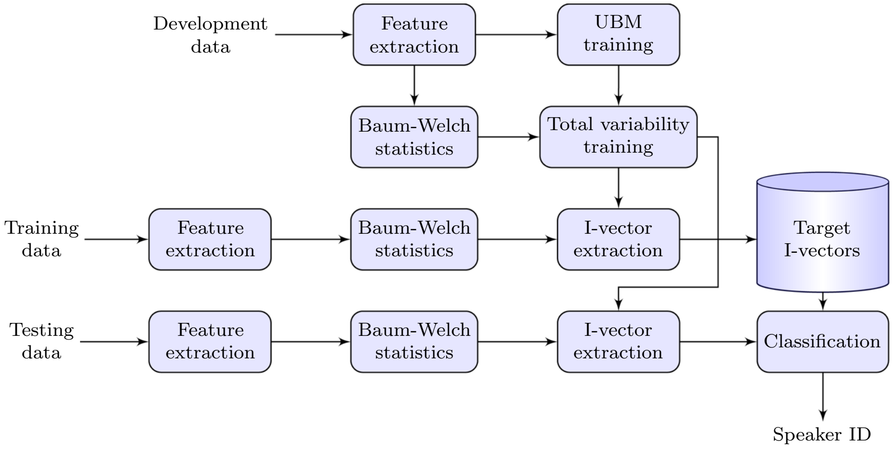

Fig. 2 The i-vector extraction process

speakers, the training data are the speech samples of known speakers, and the testing data contains speech samples of the speakers to be identified.

The details of the whole i-vector computation process is described in [12]. In short, the first step after extracting features is to compute the zeroth- and first-order sufficient statistics using a Universal Background Model (UBM). The obtained vector by stacking the mean vectors of the adapted Gaussian Mixture Model (GMM) is called supervector $\boldsymbol{s}$ and is modeled as follows [12]

$$\boldsymbol{s} = \boldsymbol{s}_{\text{ubm}} + T\text{w} \tag{1}$$

where s_{ubm} is the mean supervector coming from the UBM model, T is the total variability matrix, and w is a standard normally distributed latent variable known as the identity vector or i-vector. The total variability matrix is trained on a development set using an Expectation–Maximization (EM) algorithm and tries to capture both of speaker and session variabilities. Finally, in the testing phase, the obtained i-vectors are generally post-processed to compensate the session variability before classification.

2.2 Feature Extraction

Feature extraction is one of the key components of speaker recognition systems, and it aims at finding such a lower dimensional representation of the speech signal which would preserve speaker separability as much as possible. As shown in Fig. 3, feature characteristics can be classified into two main groups, namely physiological and behavioral characteristics [16].

Physiological characteristics reflect the physical traits of the vocal tract. They include short-term features, which are also known as low-level features. They are extracted from short frames (about 20–30 ms in duration) and detail the short-term

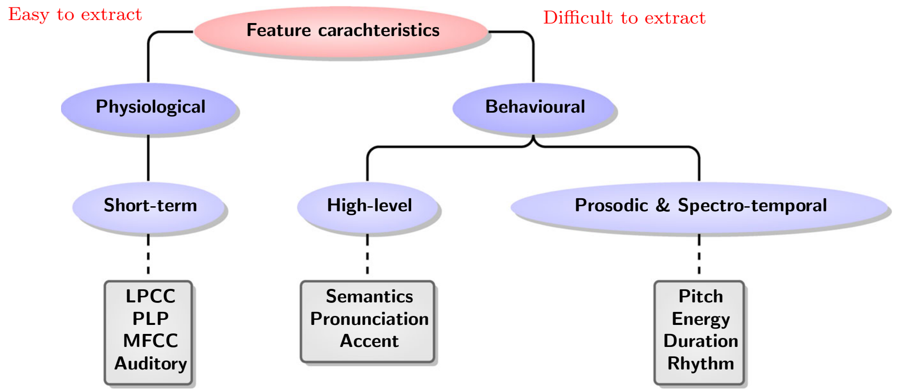

Fig. 3 Feature characteristics classification

spectral envelope which is an acoustic correlate of voice timbre. Examples of such features extractors are MFCC, Linear Predictive Coding (LPC), PLP and GFCC.

On the other hand, behavioral characteristics are based on learned speaking habits of an individual and are expected to be less susceptible to channel effects and background noise. However, they are more difficult to extract [37]. Behavioral characteristics are classified into high-level, prosodic and spectro-temporal features. Prosodic features are stress, intonation and timing measures which are mainly expressed using variations in energy, pitch and duration. Unlike low-level features, which are directly extracted from speech signals, high-level features are related to the textual content of the speech signal, including semantics, accent and pronunciation.

While all of these features appear to give useful information about speakers, their use depend primarily on the intended application. In ASR, modeling behavioral features is motivated for two main reasons. First, such features have been shown to increase performance in noisy environments when they are combined with spectral features [8, 29]. Second, they can be used not only for recognizing speakers, but also for capturing behavioral and linguistic aspects that are not reflected at the cepstral level. However, using these features alone could not give the required performance level for ASR systems [29]. Moreover, they appear to be error prone and computationally costly [6]. In this work, in order to fulfill the real-time processing requirement, we put an emphasis on the use of short-term features, especially MFCC as a trade-off between robustness, discriminability and computational complexity.

MFCCs were introduced in 1980 by Davis and Mermelstein [11]. Many steps are involved in computing MFCCs. First, the input speech signal is divided into frames of k samples and then multiplied by a smooth window function. Next, each frame is converted from time to frequency domain using FFT. In order to weight features so as to mimic human auditory perception, a Mel filter bank with a triangular band-pass frequency response is then applied to the resulting spectrum. Finally, the output of Mel filter bank undergoes logarithmic compression and DCT which converts the log

Mel spectrum back to time. In summary, MFCCs are obtained as follows

$$c_n = \sum_{m=1}^{M} \left(\log F(m) \right) \cos \left(\frac{\pi n}{M} \left(m - \frac{1}{2} \right) \right) \tag{2}$$

where $F(m)$, $1 \leq m \leq M$ is the Mel filter bank of M channels, and n is the index of the cepstral coefficient.

2.3 Classification

Classification aims at classifying a speaker into one of the pre-defined speaker classes. There are many classification approaches; all have some advantages and disadvantages at some particular use. The most popular state-of-the-art classification techniques for speaker identification are GMM [39], Artificial Neural Network (ANN) [15] and SVM [14]. The main advantages of SVM lie in their robustness and satisfactory performance over other classifiers [3,9] and thus justify our choice of using them as a classification scheme.

SVM is a discriminative classifier that attempts to construct an optimal hyperplane that separates the data according to their class labels in such a way that the separating margin between positive and negative examples is maximized [10]. The discriminant function for an input vector $\boldsymbol{x}$ is given by [50]

$$f(x) = \sum_{i=1}^{N} \boldsymbol{\lambda}_i \alpha_i K(\boldsymbol{x}, \boldsymbol{x_i}) + b \tag{3}$$

where N is the number of training instances known as support vectors, K is the kernel function, $\boldsymbol{x_i}$ are support vectors, and the constant b is the bias term. $\boldsymbol{\lambda}$ is the vector of dual variables corresponding to each separation constraint and $\alpha_i \in \{-1, +1\}$ are class labels: $+1$ for in-class and -1 for out-of-class. The selection of an efficient and effective kernel function is a key issue in applying SVM to speaker identification. Several kernel functions are defined in the literature among them linear, polynomial, multilayer perceptron (MLP) and Radial Basis Function (RBF).

The classical SVM task is a binary classification. However, to support multi-class classification problems, two basic strategies have been developed, namely one-versus-one and one-versus-all, both make use of combinations of binary classifiers. In the one-versus-all approach, each class is separated from all other classes. Thus, the number of SVM that is trained equals to the number of classes. In the one-versus-one approach, each class is separated from each other class. Here, if we consider p classes, the number of the SVM that is trained is larger and equals $p(p-1)/2$.

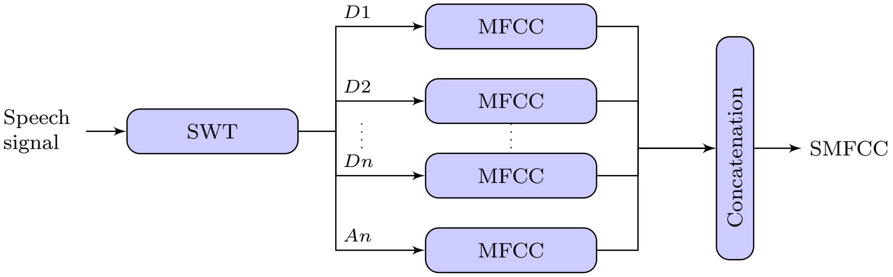

Fig. 4 Proposed feature extraction process

3 The Proposed Approach

Selecting the best feature extraction technique lies at the core of any speaker recognition system. Most of the state-of-the-art identification systems employ MFCC features. In spite of their easy deployment and generally higher accuracy, they represent some disadvantages, most significantly, their poor performance in noisy and noise-mismatched conditions. In order to improve the performance of MFCCs in noisy conditions, a series of techniques have been used including speech enhancement, feature and channel compensation and noise reduction. Therefore, applying these methods to a degraded speech is still not straightforward.

Conventionally, MFCC features are extracted over the full frequency band of the signal, which means that noise effects spread over all the feature vector components even when the noise is band limited. In this paper, we present a novel feature extraction technique that overcomes the single resolution limitation of MFCC by incorporating the time–frequency multi-resolution analysis offered by wavelet transforms. By extracting features independently for a set of timescale sub-bands, we get a large set of features available that may be able to provide complementary information, and hence, even a band-limited noise signal would not spread over the entire feature space.

The proposed feature extraction process is depicted in Fig. 4 and is summarized as following

1. First, for each speech sample, the signal is decomposed into sub-bands using SWT to achieve a series of approximation and detail coefficients, i.e., for example, for a 3-level decomposition, detail level-1 (D1), detail level-2 (D2), detail level-3 (D3) and last approximation level-3 (A3) are considered as sub-bands.
2. Then, MFCC are computed from each sub-band. Here, the first 13 coefficients derived from a 20-channel mel-scaled filterbank are extracted from speech frames of 25 ms with a frame shift of 10 ms, removing the first one because it carries less speaker-specific information [33]. In addition to static MFCC features, the log energy as well as first and second derivatives was also included to produce a feature vector of 39 elements.
3. Concatenate all the sub-band features to produce a final feature vector, denoted as SMFCC.
4. Repeat steps 1–3 for each speech sample to create a feature matrix that will be fed into the i-vector modeling framework.

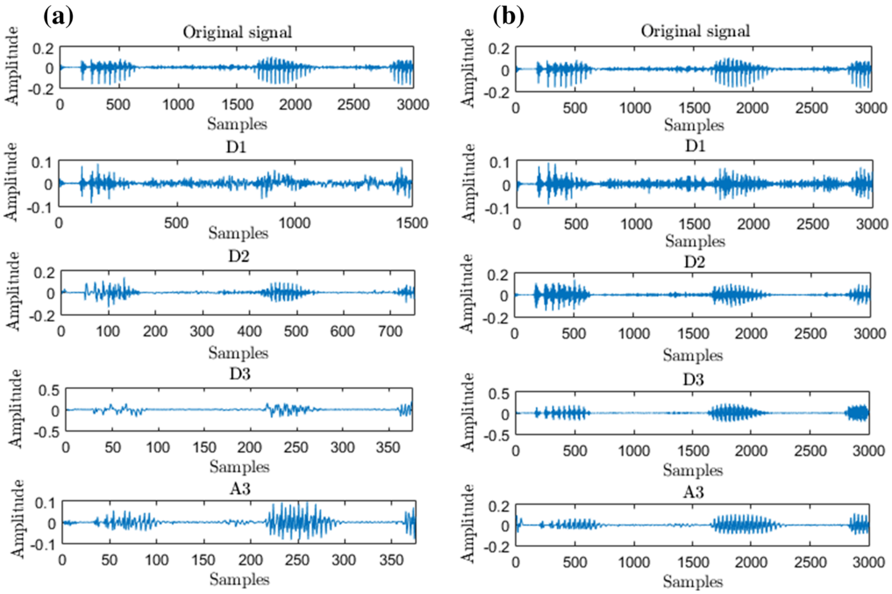

Fig. 5 Example of DWT (**a**) and SWT (**b**). Top: original signal. Bottom: 3-level decomposition using db10 wavelets. Dn represent the detail coefficients, whereas A3 represents the last approximation coefficient

The main reason for using SWT is the fact that it is a time-invariant transform as compared to DWT. As shown in Fig. 5, the size of SWT data is the same as original speech signal even after decomposition; thus, more information will be preserved.

Employing wavelet transforms brings on the challenge of selecting the appropriate mother wavelet, and choosing an improper wavelet will impact the identification accuracy. Standard wavelet families include Haar, coiflets, Daubechies and symlets. In this work, Daubechies family is used because of its effective low-pass and high-pass filter banks. It was also shown to be best suited in preserving the features of denoised signals [44].

4 Experimental Setup and Results

To test the performance of the proposed closed-set text-independent SIS, an analysis is conducted based on two datasets from two publicly available databases. The first one consists of 10 speakers (4 females and 6 males) from the Air Traffic Control Simulation (ATCOSIM) speech corpus [25]. ATCOSIM database was chosen primarily as it is an ATC operator speech corpus that was recorded during real-time simulations. This makes it suitable for being used as pristine speech data in our intended application. However, its disadvantage is the small number of speakers. The second one is a set of 455 speakers (41 females and 414 males) from Voxforge database [51], an online user-generated corpus which has the goal of collecting speech data for various languages, but only those in English were chosen for the experiments. The use of this

Table 1 Corpora used

	ATCOSIM	Voxforge
Speakers	10	455
Sampling rate (kHz)	8	8
Quality	Clean	Microphone
English level	Non-native	Mixed
Utterances per speaker	100	10

database was chosen for two main reasons. First, it is publicly available. Second, it is a larger database with more than 400 speakers. However, audio signals are recorded under uncontrolled conditions; thus, the quality of the recorded speech signals is unreliable. ATCOSIM database consists of a total of 10,078 clean utterances. In that, 100 utterances per speaker were randomly selected as an evaluation. Since the background data has to be as close as possible to the evaluation data, a subset of 300 speakers from TIMIT database [17] were used as a background set for UBM and i-vector extractor training. The speech signals were recorded using a close-talk headset microphone, with a sampling frequency of $f_s = 32$ kHz. For the experiments, they were downsampled to $f_s = 8$ kHz. On the other hand, Voxforge database consists of recordings of more than 600 speakers with a contribution of 10 utterances per volunteer speaker of approximately 4 s each. In that, recordings of 206 speakers were used to construct the background set. A brief summary of used data is given in Table 1.

In the evaluation stage, the speech files of all speakers in both databases were randomly divided into two parts. Here, 70% of clean speech samples of each speaker were used for training, and the remaining 30% were used directly or corrupted by different types of noises for testing.

4.1 Construction of Noisy Corpora

The audio recordings available from both of ATCOSIM and Voxforge databases cannot be directly used to evaluate the robustness of the proposed SIS in the presence of aeronautical noise, because they contain only clean speech signals. Hence, in order to fulfill this task, noisy signals are generated using an aeronautical communication system. It is based on OFDM, which is considered for future aeronautical communication systems, especially within the LDACS1 [41] standard, on which we focus in this paper.

Basically, an OFDM system consists of a transmitter and a receiver. As shown in Fig. 6, considering an OFDM system with N sub-carriers, the first task of the OFDM transmitter is to split the high data rate input stream to be transmitted in N lower data rate streams through a serial to parallel converter. Next, inverse fast Fourier transform (IFFT) of size N is processed on each sub-carrier data to convert it to time domain. The received signal $z(t)$ consists then of distorted versions of the transmitted OFDM symbols, which is expressed as

$$z(t) = s(t) \otimes h(t) + g(t) \tag{4}$$

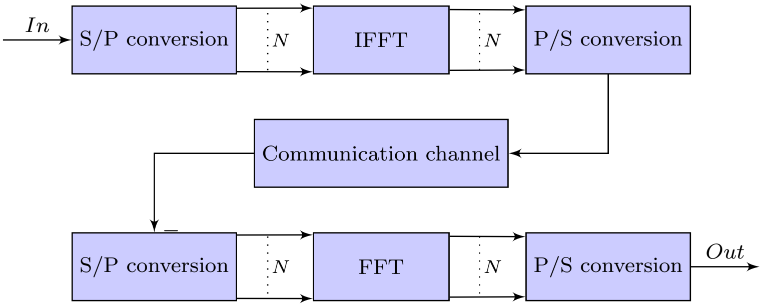

Fig. 6 Block diagram of an OFDM system

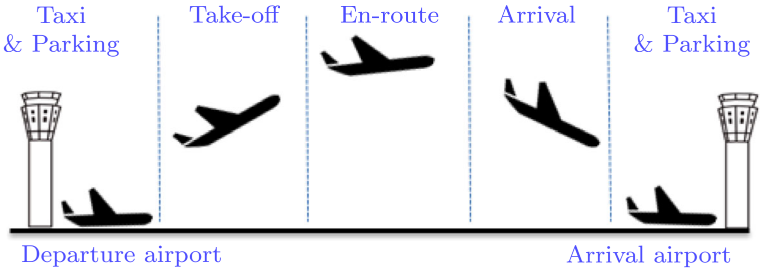

Fig. 7 Flight phases

where $\otimes$ is the cyclic convolution operation, $h(t)$ and $g(t)$ stand for the discrete time channel impulse response and the additive noise, respectively. At the receiver side, the reverse process is done so as to get the original data from the received one. The Fast Fourier Transform (FFT) is performed on each block of the N received signal samples to convert the signal back to the frequency domain. The behavior of OFDM systems is related to their performance under different channel conditions. For ATC applications, a set of realistic simulation scenarios have been proposed in the literature [20] to allow researchers assessing the performance of any developed digital mobile communication system. These are illustrated in Fig. 7 and include taxi, parking, en route, takeoff and arrival scenarios. In this paper, we concentrate on the en route one.

En route scenario describes the state where the aircraft are airborne and communicates with the control tower (air–ground communication) or with another airplane (air–air communication). Here, the channel may be characterized by a two-ray Rician channel [20], which consists of a direct path h_{LOS} as well as a reflected path h_{NLOS} each expressed as

$$h_{\text{LOS}}(t) = ae^{j\left(2\pi f_{D_{\text{LOS}}}(t-\tau_{\text{LOS}})\right)} \tag{5}$$

$$h_{\text{NLOS}}(t) = be^{j\left(2\pi f_{D_{\text{NLOS}}}(t-\tau_{\text{NLOS}})\right)} \tag{6}$$

where a and b represent the fading amplitude of the Line-Of-Sight (LOS) and diffuse paths, respectively; τ_{NLOS} and τ_{LOS} are the delays. Finally, $f_{D_{\text{LOS}}}$ and $f_{D_{\text{NLOS}}}$ are the Doppler frequencies of the LOS and diffuse paths. The LOS and reflected paths are characterized by the Rician factor κ, which is defined as the ratio of the powers of the dominant path and the diffuse components

$$\kappa = \frac{a^2}{b^2} \tag{7}$$

or equivalently in decibels

$$\kappa = 10 \log_{10}(a^2/b^2) \tag{8}$$

For a simple implementation of the multipath fading channel models, it is required that the mean throughput power remains unchanged, i.e., $a^2 + b^2 = 1$. Then, the amplitude of the LOS and diffuse components are obtained from Eq. (8) as a function of the Rician factor as follows

$$a = \sqrt{\frac{\kappa}{\kappa + 1}} \tag{9}$$

$$b = \sqrt{\frac{1}{\kappa + 1}} \tag{10}$$

h_{NLOS} is a reflected and delayed path that experiences Rayleigh fading. Assuming that all processes are wide sense stationary, the envelope of the h_{NLOS} path is represented by

$$\begin{aligned} r(t) = 2 \sum_{n=1}^{M} \cos(\beta_n) \cos(\omega_n t) + \sqrt{2} \cos(\alpha) \cos(\omega t) \\ + 2 \sum_{n=1}^{M} \sin(\beta_n) \cos(\omega_n t) + \sqrt{2} \sin(\alpha) \cos(\omega t) \end{aligned} \tag{11}$$

with N is the number of multipath components, $M = (N/2 - 1)/2$ represents the number of complex sinusoids to be generated with the frequencies of $\omega_n = \omega \cos(\frac{2\pi n}{N})$, having $\omega = 2\pi f$. $\beta_n = \pi n/N$ and $\alpha = \pi/4$ are the initial phases of the nth Doppler-shifted sinusoid and the maximum Doppler frequency f, respectively. In this paper, our emphasis is on the characterization of air–ground en route communications. In this context, clean testing signals were transmitted through an OFDM system with a 128-point FFT and 32 sub-carriers. The following transmission channels were simulated:

- Ideal channel where no alteration is experienced.
- AWGN with a SNR of 5 dB, 10 dB and 15 dB.
- En route channel with a Rician factor of $\kappa = 15$ dB, which is the typical value in the en route scenario [20].
- A combination of en route and AWGN, each as described above.

4.2 Experiments

To quantify the accuracy of the proposed SIS, a comparative analysis is conducted based on the Identification Rate (IR), which is computed by

$$\text{IR}(\%) = \frac{\text{number of correctly identified utterances}}{\text{total number of utterances under test}} \times 100 \tag{12}$$

Table 2 Identification rates (%) of I-SMFCC features for ATCOSIM database

	Decomposition level									
	1	2	3	4	5	6	7	8	9	10
Wavelet order										
db1	96	95.67	97.33	95.33	94.67	96.67	97	96.67	98	96.67
db2	97	95.33	96.33	95.67	95.33	96.33	97	97.67	97	98.67
db3	94	95	95.33	95.33	95	96.33	96.67	97	94.33	94.33
db4	95.67	95.67	96.67	94.33	93.33	96	93.67	95.33	95.33	93.67
db5	97.67	96	94.67	94.67	93.33	92	95.67	93.33	93.67	91.67
db6	93.33	95.67	96.33	94	94.33	93.67	93.33	93.33	91.67	92.67
db7	94	96.33	94.33	93	94.67	95	93.33	92.33	91.33	94
db8	96	94.33	97	94	97	92.67	94.67	94.67	93.67	93
db9	94.33	96	92.67	94.33	97	92.67	94.67	94.67	93.67	93
db10	94	95.33	94	95.33	95	96.67	94.33	91	91.33	92.33

In order to achieve the most effective denoising with SWT, an investigation of the optimum parameters is conducted. These include the number of decomposition levels and the mother wavelet order.

4.2.1 Choice of Wavelet Mother Order and Decomposition Level

In this experiment, we evaluate the effect of decomposition level that is used in the performance of SMFCC features. The enrollment speech signals were kept in clean conditions, while testing speech signals were corrupted by an AWGN at a SNR of 10 dB. Both of the training and testing speech signals were subjected to ten decomposition levels using Daubechies family wavelets. MFCC features were extracted from each of the detail and the last approximation coefficients. The resulting feature vectors are then assembled, modeled using the i-vector approach and fed to the SVM classifier with a linear kernel. Tables 2 and 3 show the obtained IRs for both ATCOSIM and Voxforge databases. Regarding the performance of SMFCC features in the presence of AWGN at 10 dB, our results clearly show that when using Voxforge, the accuracy of the SIS shows a trend of increase with more wavelet decomposition levels till 9 levels, and a decrease or no improvement afterward. Thus, level 9 seems to be adequate and is used in the next experiments. For ATCOSIM database, the highest accuracy is obtained using the tenth level of decomposition, and hence, level 10 is chosen as optimum in the next experiments.

The choice of the mother wavelet is a key issue in wavelet analysis influencing the overall performance of the proposed system. The second goal of this experiment is to find the adequate mother wavelet between the ten first members of Daubechies family for wavelet-based features. Based on these findings, the best performing methods for both databases are selected in the next experiments. Tables 2 and 3 show how identification rates change, respectively, to the wavelet choice between the ten first members of Daubechies family for SMFCC features. Obviously, the wavelet that

Table 3 Identification rates (%) of I-SMFCC features for Voxforge database

	Decomposition level									
	1	2	3	4	5	6	7	8	9	10
Wavelet order										
db1	84.03	83.59	86.30	87.77	89.38	88.50	89.60	89.74	90.99	90.62
db2	83.59	85.86	86.67	87.55	89.01	89.30	90.04	90.40	91.28	91.21
db3	83.88	85.20	85.49	87.99	89.60	90.40	89.74	89.74	90.70	90.11
db4	81.98	84.03	84.47	85.86	87.47	89.45	89.45	90.55	90.70	90.48
db5	81.32	82.12	83.74	85.93	87.55	89.23	89.89	89.23	89.67	90.11
db6	80.07	81.76	83.37	85.49	86.89	88.57	88.57	88.64	89.30	88.64
db7	78.46	79.49	83.30	85.64	86.01	87.33	65.27	88.42	94.29	88.79
db8	79.49	80.29	82.71	83.15	86.08	87.77	87.33	87.62	87.55	87.84
db9	79.27	80	81.69	84.84	86.23	87.32	87.11	87.18	88.06	87.91
db10	78.53	80.95	82.27	83.08	85.86	85.93	85.79	91.50	86.67	87.18

maximizes the identification rates is chosen as optimum, “db2” for ATCOSIM and “db7” for Voxforge in this case.

4.2.2 Overall System Evaluation

In this paper, we establish a robust text-independent closed-set speaker identification system based on a multi-resolution feature extraction method.

In order to verify the accuracy of the presented method, a series of speaker identification experiments was performed. The performance analysis is conducted from two perspectives, i.e., clean and noisy environments. Furthermore, a performance comparison with and without using i-vector modeling was conducted.

For i-vector extraction, we have extracted 400-dimensional i-vectors, and the number of UBM components is 512 as in [28]. Session compensation with the i-vector approach was considered in using Linear Discriminant Analysis (LDA) and Within-Class Covariance Normalization (WCCN). LDA is applied in order to reduce the dimensionality based on the Fisher’s criterion that in turn will help to minimize within-class variability and maximize between-class variability. WCCN is used to reduce the variability, and thus, these two methods together help to increase the overall identification accuracy of the system. On the other hand, without using the i-vector modeling, 39-dimensional MFCC features were extracted in the same manner as described in Sect. 3. Then, the mean of the resulting features was selected to build the basic MFCC feature set that is then input to SVM. For SMFCC features, the same experiments were conducted as for I-SMCC features to find the adequate mother wavelet as well as the decomposition level. Hence, eight levels of decomposition were chosen using db6. Afterward, the mean of the extracted MFCC coefficients at each sub-band was computed, and all the mean vectors were concatenated to produce the final feature vector that will be fed to SVM after applying LDA and WCCN. The computed features, with

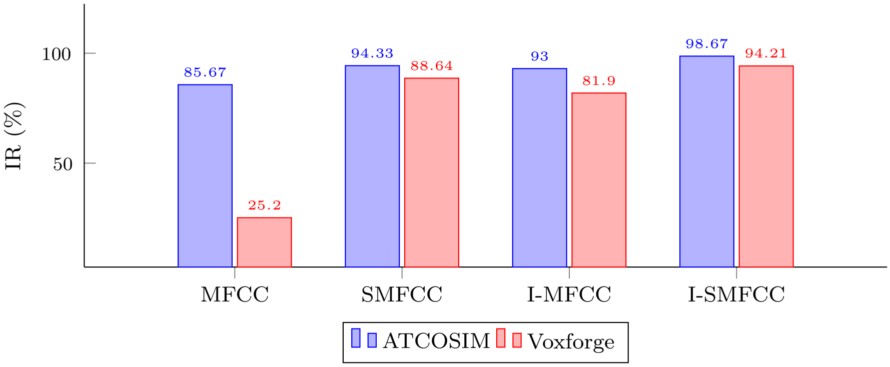

Fig. 8 Accuracy of different extraction processes in clean environment

Table 4 Accuracy (%) of the proposed method in noisy environments using ATCOSIM speech corpus

Transmission channel	SNR (dB)	MFCC	SMFCC	I-MFCC	I-SMFCC
AWGN	5	49.33	24	19.67	19.33
	10	85.67	94.33	92.67	98.67
	15	85.67	94.33	92.67	98
En route	Infinite	85.67	94.33	92.67	98
En route + AWGN	5	32	20.67	12	15
	10	84	87	91	94.67
	15	85.33	94.33	92.67	98

and without i-vector modeling, were further normalized to a unit norm using z-score normalization.

In the first set of the experiments, we examined the performance of the system in a clean environment. Figure 8 shows the comparison identification accuracy between the proposed method with and without wavelet analysis and also with and without i-vector modeling. The extracted features based on the i-vector approach are denoted as I-MFCC and I-SMFCC. We therefore conclude that globally, SMFCC improves the IR for both databases compared to baseline MFCC, with an improvement of about 6% for ATCOSIM using the i-vector approach and 23% for Voxforge without i-vector modeling.

The next experiments aim at exploring the benefit of extracting MFCC features from wavelet features to perform speaker identification over an OFDM-based communication system. A variety of different transmission channels were simulated. The SNR-dependent identification accuracies for all the evaluated noisy conditions are presented for both databases. The obtained results are presented in Tables 4 and 5.

Table 4 relates the obtained results using ATCOSIM database. We can see that in general, I-MFCC and I-SMFCC outperform MFCC and SMFCC except at 5 dB case where the performance of i-vector modeling approach is poorer. On the other hand,

Table 5 Accuracy (%) of the proposed method in noisy environments using Voxforge database

Transmission channel	SNR (dB)	MFCC	SMFCC	I-MFCC	I-SMFCC
AWGN	5	1.54	7.99	3.44	6.15
	10	25.20	88.79	81.83	94.29
	15	25.20	88.64	81.83	94.21
En route	Infinite	25.20	88.64	81.83	94.21
En route + AWGN	5	1.10	5.05	1.47	3.44
	10	22.05	82.56	79.05	87.03
	15	25.35	88.64	81.90	94.21

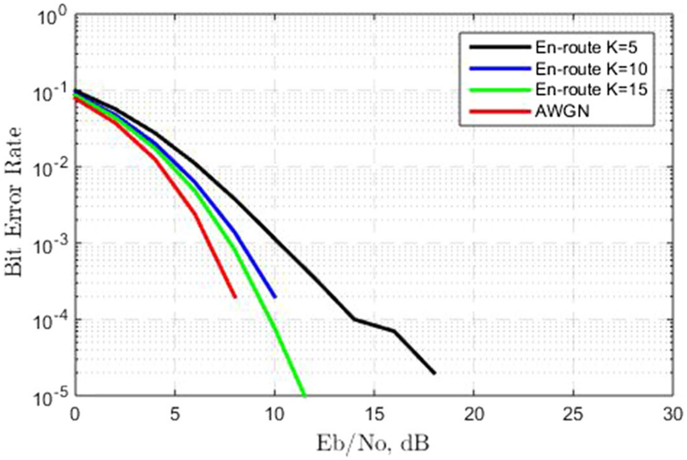

Fig. 9 BER performance of OFDM over AWGN and en route channels for different Rician κ values

Table 5 reports the results obtained using Voxforge database, where the advantage of using SMFCC is again obvious compared to baseline MFCC. Except for 5 dB, although there is an improvement compared to MFCC, the achieved performance remains poor.

To get a deeper understanding of the performance in noisy environments and especially in case of 5 dB, we report the Bit Error Rate (BER) analysis of OFDM communication system in Fig. 9 that was also investigated in [42]. The reported results indicate that at 5 dB, the BER is around 10^{-2}. This very high value makes the possibility of detecting the correct signal less likely. While at 10 dB, the BER is around 10^{-5}. Since features are extracted directly from speech signals received from OFDM system, this justifies the different behaviors of the SIS at 5 dB and 10 dB.

5 Conclusion

In this paper, wavelet-based features have been examined for the purpose of integrating ASR in future aeronautical communication system where MFCC coefficients were extracted from SWT sub-bands. Two approaches were compared, with and without i-vector modeling approach. A proof-of-concept implementation was provided by considering several transmission channels. By using wavelet analysis, the proposed method was demonstrated to be capable of providing satisfactory performance in the en route noisy environment. At SNRs higher than 10 dB, the advantage of the SWT-based features in terms of identification accuracy is larger than 13%. Speaker identification experiments showed that the usage of only one statistic, when modeling without the i-vector representation, yields to a satisfactory performance of the overall system for ATCOSIM database. As part of future work, we aim to investigate more robust features for extracting speaker characteristics at lower SNRs.

References

1. S. Abd El-Moneim, M.I. Dessouky, F.E. Abd El-Samie, M.A. Nassar, M. Abd El-Naby, Hybrid speech enhancement with empirical mode decomposition and spectral subtraction for efficient speaker identification. Int. J. Speech Technol. **18**(4), 555–564 (2015)
2. N. Asbai, A. Amrouche, M. Debyeche, Performances evaluation of GMM-UBM and GMM-SVM for speaker recognition in realistic world, in *Neural Inf. Process.*, ed. by B.-L. Lu, L. Zhang, J. Kwok (Springer, Berlin, 2011), pp. 284–291
3. J.G.P. Bernal, A.P. Guerrero, J.G. Close, A speaker verification system using SVM over a Spanish corpus, in *2009 Mexican International Conference on Computer Science*, Sept 2009, pp. 381–386
4. J.P. Campbell, Speaker recognition: a tutorial. Proc. IEEE **85**(9), 1437–1462 (1997)
5. J.P. Campbell, D.A. Reynolds, Corpora for the evaluation of speaker recognition systems, in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 829–832, March 1999
6. M. Carey, E. Parris, H. Lloyd-Thomas, S. Bennett, Robust prosodic features for speaker identification, in *Proceedings of ICSLP-96*, November 1996
7. P.M. Chauhan, N.P. Desai, Mel Frequency Cepstral Coefficients (MFCCs) based speaker identification in noisy environment using wiener filter, in *2014 International Conference on Green Computing Communication and Electrical Engineering (ICGCCEE)*, Mar 2014, pp. 1–5
8. S.-H. Chen, H.-C. Wang, Improvement of speaker recognition by combining residual and prosodic features with acoustic features, in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, May 2004, pp. 93–96
9. B.J. Chua, X.J. Li, H.D. Tran, Study of automatic biosounds detection and classification using SVM and GMM, in *2011 IEEE/NIH Life Science Systems and Applications Workshop (LiSSA)*, April 2011, pp. 155–158
10. C. Cortes, V. Vapnik, Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
11. S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust., Speech Signal Process. **28**(4), 357–366 (1980)
12. N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front-end factor analysis for speaker verification. IEEE Trans. Acoust., Speech Signal Process. **19**(4), 788–798 (2011)
13. S. Dey, S. Barman, R.K. Bhukya, R.K. Das, B.C. Haris, S.R.M. Prasanna, R. Sinha, Speech biometric based attendance system, in *2014 Twentieth National Conference on Communications (NCC)*, Feb 2014, pp. 1–6
14. H. Ding, Z.-M. Tang, L.-H. Wei, Y.-P. Li, A study on speaker identification based on weighted LS-SVM. Autom. Control Comput. Sci. **43**(6), 328–335 (2009)
15. M. Dutta, C. Patgiri, M. Sarma, K.K. Sarma, *Closed-Set Text-Independent Speaker Identification System Using Multiple ANN Classifiers* (Springer, Cham, 2015), pp. 377–385
16. M. Faundez-Zanuy, E. Monte-Moreno, State-of-the-art in speaker recognition. IEEE Aerosp. Electron. Syst. Mag. **20**(5), 7–12 (2005)
17. J. Garofolo, L. Lamel, Fisher, J. Fiscus, D. Pallett, N. Dahlgren, V. Zue, Timit acoustic-phonetic continuous speech corpus. *Linguistic Data Consortium* (1993)

18. H. Gish, M. Schmidt, Text-independent speaker identification. IEEE Signal Process. Mag. **11**(4), 18–32 (1994)
19. S.M. Govindan, P. Duraisamy, X. Yuan, Adaptive wavelet shrinkage for noise robust speaker recognition. Digit. Signal Process. **33**, 180–190 (2014)
20. E. Haas, Aeronautical channel modeling. IEEE Trans. Veh. Technol. **51**(2), 254–264 (2002)
21. M. Hagmüller, G. Kübin, Speech watermarking for air traffic control. Eurocontrol Experimental Centre, EEC Note 05/05 (2005)
22. M. Hébert, *Text-Dependent Speaker Recognition* (Springer, Berlin, 2008), pp. 743–762
23. HINDSIGHT, Nn°2 communication. Technical report, EUROCONTROL (2006)
24. K. Hofbauer, H. Hering, G. Kübin, Speech watermarking for the VHF radio channel, in *4th EUROCONTROL Innovative Research Workshop*, December 2005
25. K. Hofbauer, S. Petrik, H. Hering, The ATCOSIM corpus of non-prompted clean air traffic control speech, in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008
26. Y. Hu, P. Loizou, Subjective comparison and evaluation of speech enhancement algorithms. Speech Commun. **49**, 588–601 (2007)
27. A.A.A.A. Khalil, E.S.M. Saad, M.A. El-Nabi, F.E.A. El-Samie, Efficient speaker identification from speech transmitted over Bluetooth based system, in *2013 8th International Conference on Computer Engineering Systems (ICCES)*, Nov 2013, pp. 190–193
28. W.B. Kheder, D. Matrouf, P.-M. Bousquet, J.-F. Bonastre, M. Ajili, Fast i-vector denoising using map estimation and a noise distributions database for robust speaker recognition. Comput. Speech Lang. **45**, 104–122 (2017)
29. S.G. Koolagudi, K. Sreenivasa Rao, R. Reddy, V.A. Kumar, S. Chakrabarti, *Robust Speaker Recognition in Noisy Environments: Using Dynamics of Speaker-Specific Prosody* (Springer, New York, 2012), pp. 183–204
30. K.A. Lee, A. Larcher, H. Thai, B. Ma, H. Li, Joint application of speech and speaker recognition for automation and security in smart home, in *INTERSPEECH*, 2011, pp. 3317–3318
31. K.A. Lee, B. Ma, H. Li, Speaker verification makes its debut in smartphone. *IEEE Signal Processing Society Speech and Language Technical Committee Newsletter* (2013)
32. Y. Lei, L. Burget, N. Scheffer, A noise robust i-vector extractor using vector Taylor series for speaker recognition, in*2013 IEEE international conference on acoustics, speech and signal processing*, May 2013, pp. 6788–6791
33. B.G. Nagaraja, H.S. Jayanna, *Multilingual Speaker Identification with the Constraint of Limited Data Using Multitaper MFCC* (Springer, Berlin, 2012), pp. 127–134
34. M. Neffe, V. Pham, H. Horst, G. Kubin, Speaker segmentation for air traffic control, in *Lecture Notes in Artificial Intelligence* (Springer, Berlin, 2007), pp. 177–191
35. NIST. The NIST year 2002 speaker recognition evaluation plan. National Institute of Standards and Technology of USA, February (2002). Available: http://www.nist.gov/speech/tests/spk/2002/doc/2002-spkrec-evalplan-v60.pdf
36. S. Qureshi, I. Masood, M. Hashmi, S. Hanninen, M. Sarwar, A. Jameel, Noise reduction of electrocardiographic signals using wavelet transforms. Elektron. Elektrotech. **20**(3), 29–32 (2014)
37. D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, B. Xiang, The Supersid Project: exploiting high-level information for high-accuracy speaker recognition, in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)*, vol. 4, Apr 2003, pp. IV–784
38. D.A. Reynolds, An overview of automatic speaker recognition technology, in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, May 2002, pp. IV–4072–IV–4075
39. D.A. Reynolds, R.C. Rose, Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. Speech Audio Process. **3**(1), 72–83 (1995)
40. S. Sadjadi, J. Hansen, Mean Hilbert envelope coefficients (MHEC) for robust speaker and language identification. Speech Commun. **72**, 138–148 (2015)
41. M. Sajatovic, et al., L-DACS1 system definition proposal: ddeliverable D2. Technical report version 1.0, Feb 2009
42. S. Sekkate, M. Khalil, A. Adib, An improved automatic aircraft identification system, in *2016 International Conference on Wireless Networks and Mobile Communications (WINCOM)*, Oct 2016, pp. 47–51

43. S. Sekkate, M. Khalil, A. Adib, Speaker identification: a way to reduce call-sign confusion events, in *2017 International Conference on Advanced Technologies for Signal & Image Processing*, May 2017
44. U. Seljuq, F. Himayun, H. Rasheed, Selection of an optimal mother wavelet basis function for ECG signal denoising, in *17th IEEE International Multi Topic Conference 2014*, Dec 2014, pp. 26–30
45. S. Selva Nidhyananthan, R. Shantha Selva Kumari, T. Senthur Selvi, Noise robust speaker identification using RASTA–MFCC feature with quadrilateral filter bank structure. Wirel. Pers. Commun. **91**(3), 1321–1333 (2016)
46. A. Shafik, S.M. Elhalafawy, S.E.M. Diab, B.M. Sallam, F.E.A. El-Samie, A wavelet based approach for speaker identification from degraded speech. IJCNIS **1**(3), 52–58 (2009)
47. Y. Shao, S. Srinivasan, D. Wang, Incorporating auditory feature uncertainties in robust speaker identification, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2007*, Honolulu, HI, USA, 15–20 Apr 2007, pp. 277–280
48. D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, X-vectors: robust DNN embeddings for speaker recognition, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018), pp. 5329–5333
49. M. Stark, T.V. Pham, F. Pernkopf, G. Kubin, H. Hering, Speaker verification for air traffic control, in *EUROCONTROL Innovative Research Workshop and Exhibition*, Dec 2006
50. V.N. Vapnik, *Statistical Learning Theory. Adaptive and Learning Systems for Signal Processing, Communications, and Control* (Wiley, New York, 1998)
51. Voxforge database. Technical report
52. A. Vuppala, K.S. Rao, Speaker identification under background noise using features extracted from steady vowel regions. Int. J. Adapt Control Signal Process **27**(9), 781–792 (2013)
53. L. Xu, R.K. Das, E. Yilmaz, J. Yang, H. Li, Generative x-vectors for text-independent speaker verification (2018). *CoRR*, arXiv:1809.06798
54. X. Zhao, Y. Shao, D. Wang, Casa-based robust speaker identification. IEEE Trans. Audio Speech Lang. Process. **20**(5), 1608–1616 (2012)
55. X. Zhao, Y. Wang, D. Wang, Robust speaker identification in noisy and reverberant conditions. IEEE Trans. Audio Speech Lang. Process. **22**(4), 836–845 (2014)
56. T.F. Zheng, Q. Jin, L. Li, J. Wang, F. Bie, An overview of robustness related issues in speaker recognition, in*Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, Dec 2014, pp. 1–10

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.