SIViP (2014) 8:365–375
DOI 10.1007/s11760-012-0299-z

ORIGINAL PAPER

Feature classification criterion for missing features mask estimation in robust speaker recognition

Dayana Ribas González · José Ramón Calvo de Lara

Received: 15 December 2010 / Revised: 22 February 2012 / Accepted: 27 February 2012 / Published online: 20 March 2012

Abstract Currently, many speaker recognition applications must handle speech corrupted by environmental additive noise without having a priori knowledge about the characteristics of noise. Some previous works in speaker recognition have used the missing feature (MF) approach to compensate for noise. In most of those applications, the spectral reliability decision step is performed using the signal to noise ratio (SNR) criterion, which attempts to directly measure the relative signal to noise energy at each frequency. An alternative approach to spectral data reliability has been used with some success in the MF approach to speech recognition. Here, we compare the use of this new criterion with the SNR criterion for MF mask estimation in speaker recognition. The new reliability decision is based on the extraction and analysis of several spectro-temporal features from across the entire speech frame, but not across the time, which highlight the differences between spectral regions dominated by speech and by noise. We call it the feature classification (FC) criterion. It uses several spectral features to establish spectrogram reliability unlike SNR criterion that relies only in one feature: SNR. We evaluated our proposal through speaker verification experiments, in Ahumada speech database corrupted by different types of noise at various SNR levels. Experiments demonstrated that the FC criterion achieves considerably better recognition accuracy than the SNR criterion in the speaker verification tasks tested.

Keywords Speaker verification · Missing feature approach · Mask estimation

D. Ribas González (✉) · J. R. Calvo de Lara
Advanced Technologies Application Center (CENATAV),
7a ave. 21812 Siboney, Playa, 12200 Havana, Cuba
e-mail: dribas@cenatav.co.cu

J. R. Calvo de Lara
e-mail: jcalvo@cenatav.co.cu

1 Introduction

Nowadays, automatic speaker recognition is widely used in biometric applications like remote authentication, forensic research, detection and tracking of speakers. Usually, these applications work in uncontrolled environments, so using single channel speech signals acquired in noisy acoustic environments, such as telephone booths, hidden microphones, mobile phones and multispeaker environments, is very common. In these cases, noise is added to the speech signal causing bad performance. In order to handle environmental additive noise, many compensation techniques applied directly to the signal (speech enhancement methods) or to some system stage have been proposed, most of them for speech recognition applications.

Examples of speech enhancement methods used in speaker recognition systems are the well-known filtering techniques, Wiener filtering [1] or Spectral Subtraction [2]. They assume a priori knowledge of the noise spectrum, and therefore, they frequently use noise estimation techniques [3,4]. On the other hand, other methods could be applied in each stage of speaker recognition application: parameterization, modeling, comparison (score's computation), and in the train-test matching conditions, known as multicondition training method [5]. In parameterization, speaker features representations, more robust to noise than others have been developed, like MFCC [6] and PLP [7]; however, these methods are not robust enough to obtain high accuracy in speaker recognition using highly corrupted speech signals.

In modeling, there are compensation techniques based on the integration of model noise spectrum to speaker model, examples of these are parallel model combination (PMC) [8] and Jacobian environmental adaptation [9]. These methods assume that the characteristics of noise are previously known but this fact cannot be possible in real life applications. One could think that noise estimation is the solution to this problem, but if it is obtained with an unreliable estimator the system performance could degrade a lot. Besides most acoustic additive noises that appear in real scenarios are very hard to estimate, for example, mixture of noises, non-stationary noises, and noises correlated with speech.

In comparison, score normalization methods [10] like Z-norm, T-norm, could be used to deal with noise. This category of methods is very useful to cope with score variability and requires relatively little a priori knowledge of noise characteristics. However, to obtain really good normalization parameters, an impostor voice set must be acquired in the same conditions of the target voice set, and no database meets this condition. Furthermore, these methods are highly data-driven and require a lot of data for training cohort models and it is not trivial to decide how to split the corpus for score normalization [11]. As seen previously, it is difficult to obtain the necessary data to perform an adequate score normalization in real applications.

In order to overcome the limitations of noise compensation techniques, the missing feature (MF) approach [12] has been applied. Unlike others, MF was designed to handle unknown noise and does not require a priori knowledge of corrupted noise characteristics. The MF paradigm is based on the fact that any noise affects time–frequency (t–f) regions of the speech spectrum in different ways and it consists in detecting spectrum corruption level and determining which part of the spectrum is reliable enough to be used in recognition.

Use of the MF approach in speech processing has two steps. The first is missing feature detection, which consists in the detection of the reliability degree of the corrupted speech spectrum, by creating a map of the reliability in each t–f region, called a spectrographic mask. The mask is formed by reliable (R) and unreliable (U) labels for each t–f region in the spectrum. Regions highly corrupted by noise are tagged with U labels and the regions with a low level of corruption with R labels. The second step is missing feature compensation, based on the spectrographic mask. This has two options: to reconstruct unreliable regions to perform recognition with the newly reconstructed spectrum or to bypass unreliable regions, so as not to use it in the recognition process. The first option uses reconstruction or missing data imputation techniques, developed for speech recognition [13]. The second is known as marginalization and requires a change in score computation method to handle an incomplete set of spectral features in speaker verification. In [14] was shown the better performance of marginalization over imputation techniques. Other kinds of marginalization were developed first for speech recognition [34] and later for speaker recognition, such as bounded marginalization [15] and accurate marginalization [16].

Published results [17–19] show that the MF method is capable of providing robustness to speaker recognition in noisy environment, however, while the potential for improvement increases, it is mainly dependent on mask estimation accuracy. This happens because missing feature compensation works only with unreliable regions determined by mask. If the mask is not accurate, some errors will be introduced, that is, some reliable regions will be damaged, while some unreliable ones will be unchanged. In short, it could be said that mask estimation is the main process in the MF approach, so in this article, we will focus on the mask estimation step.

The most frequently used criterion to estimate the mask in speaker recognition is SNR[1] [20], which consists in computing local SNR in each t–f region of the speech spectrum for determining the reliability measure to be used in speaker recognition. This paradigm, that we will call: “SNR criterion”, uses various methods to compute local SNR [17,19,21,22]. However, computing SNR accurately is hard, especially when the signal is corrupted by some kind of noise other than stationary. These methods rely only on SNR computation, so if this is not accurate enough the spectrographic mask estimation will not be either.

The goal of this study is to evaluate the accuracy of speaker recognition in noisy speech using a method for estimating spectrographic mask for MF approach following the paradigm of evaluating different spectro-temporal information for determining the spectrum reliability. This paradigm, that we will call the “Features Classification (FC) criterion”, improves over the SNR criterion making use of several complementary features, hence if one is affected by noise the others could ensure that reliability decision remains consistent. For that purpose, the method proposed by Seltzer et al. [23] originally applied in robust speech recognition has been used. Roughly, this method divides the speech signal into t–f spectral regions using a Mel Filterbank, extracts several spectro-temporal features in those regions which enhance the differences between t–f regions dominated by noise and t–f regions dominated by speech, then it uses a binary Bayesian classifier to determine the reliable and unreliable spectral t–f regions to be used in the speaker recognition task. For unreliable compensation, we used marginalization of the unreliable spectrum [14,17]. To evaluate the robustness of the FC criterion in speaker recognition, we conducted speaker verification experiments in several noisy environments, applying the MF approach with the FC criterion and compared it with the results obtained applying SNR criterion.

From now on, this article is organized as follows. Section 2 explains the MF schema proposed and the mask estimation methods used as baseline. Section 3 presents the experimental setup of speaker verification tests. Section 4

[1] SNR: Signal to noise ratio is a measure to quantify how much a signal has been corrupted by noise. It is defined as the power ratio between a signal and the background noise.

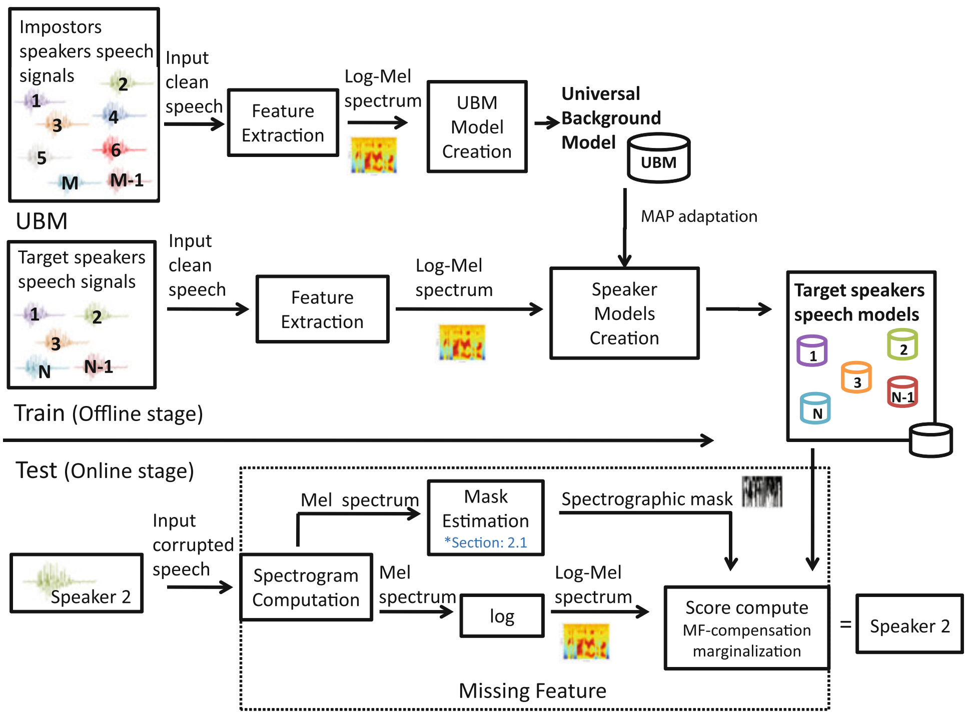


Fig. 1 Speaker recognition system based on missing feature approach

explains the results and discussions. Conclusions and future work are referred to in Sect. 5.

2 Missing feature schema proposed

Figure 1 shows a diagram of a speaker recognition system based on the MF approach, in general the speaker recognition system follows the GMM-UBM-MAP state of the art paradigm proposed by Reynolds et al. [24]. The system is trained with clean speech signals to obtain speaker models and MF techniques are applied only to test corrupted speech signals, where the system first computes the Mel spectrum of input corrupted speech and later, in order to determine its reliability, a mask estimator is applied labeling R and U spectral regions. Finally, Mel spectral features are taken in logarithmic scale and a marginalized score is computed using only the R regions determined by mask, ignoring U regions.

2.1 Mask estimator based on FC criterion

Figure 2 shows a diagram of the mask estimation method [23] used to apply FC criterion in speaker recognition, it is taking the place of the mask estimation block in the system presented previously (Fig. 1).

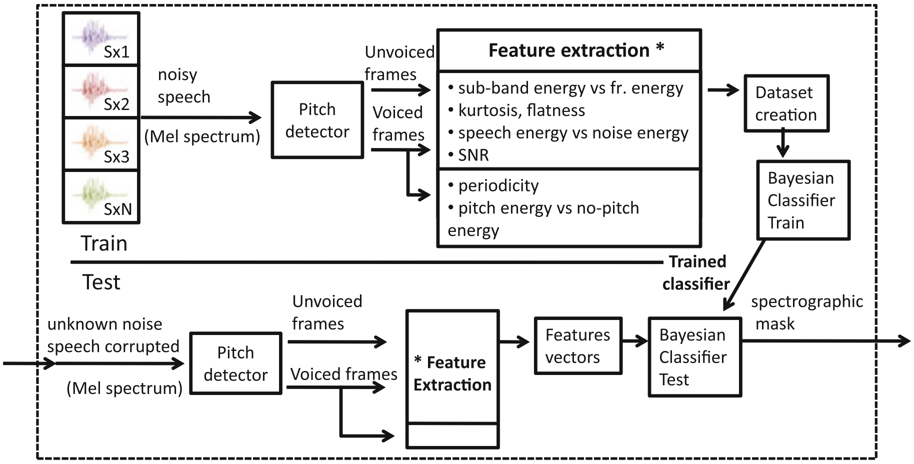
For determining spectrum reliability this method relies on several independent features extracted from each t–f region of speech spectrum—called mask features—and determined by 20 Mel filterbanks applied to each frame. As noise affects voiced and unvoiced frames in a different way, a robust algorithm for pitch tracking [25] was implemented and applied to divide voiced speech frames from unvoiced speech frames. Specifically, seven mask features were extracted from voiced frames and only five from unvoiced frames, because there are the two that depend on pitch can only be used in voiced frames.

Those features were introduced into a supervised Bayesian classifier scheme—called mask classifier—, which was in charge of determining the spectrum reliability by classifying each t–f region as U or R class, depending on its level of corruption. The training scheme was multicondition, using four types of noise: stationary pink noise, pseudo-stationary exhibition noise, music noise (theme of Pulp Fiction film), and non-stationary restaurant. The mask classifier was trained with a lot of mask feature samples extracted from 25,179 speech signals corrupted at SNR levels from 0 to 25 db. An Oracle mask[2] was used for labeling (R/U) the training dataset. The a priori probabilities of R and U classes are taken as the relative frequency of each class in the training set.

Later on, the number of classifiers to create was fixed, taking into account the following arguments. Firstly, since the number of features used for voiced speech segments was different from unvoiced speech segments, the decision was to use one classifier for voiced segments and another for unvoiced segments. Secondly, the spectrum behavior and

[2] Oracle mask: is the ideal spectrographic mask, which is computed using truly local SNR in each t–f component, given both the clean and noisy speech signals, this mask has often been used to establish upper-bound limits on recognition accuracy that can be obtained using MF.

Fig. 2 Mask estimation method based on FC criterion



the values of the mask features obtained for each subband was analyzed. Then, it was observed that the values of mask features in a single frame changed considerably across frequency subbands, less for stationary noise. This happens because of the different manifestation of frequencies in a voice, specially in subbands that contain speech formants. So, finally a separate classifier was trained for voiced and unvoiced segments, and for each of the 20 subbands of the Mel Filterbank used in the feature extraction, for a total of 40 classifiers.

Our own implementation of this method was used in this article, where the classification scheme was designed and implemented supported by the PRTool toolkit [26].

2.1.1 *Mask features analysis*

The original motivation of Seltzer et al. [23] for this method was to apply it on a speech recognition task. Since our aim is for speaker recognition, and in order to decide which mask features would be selected for our proposal, in this section an analysis of the particular contribution of each mask feature proposed in speaker recognition was done.

The first mask feature was computed as the log ratio of energy in a subband regarding the energy in the overall corresponding frame. This feature measures the contribution of subband frequency components in the whole frame spectral energy. Clean speech has most of its spectral energy in low frequencies and in voiced frames, white additive noise tends to increase spectral energy of frequencies where it is manifested, provoking changes in the utterance spectral energy distribution as a function of noise frequencies and SNR level. So, when the mask classifier is trained from clean to very noisy speech, the supervised labeling for clean speech is done fixing as R those regions with relatively high energy in low frequencies and relatively low energy high frequencies for voiced frames, while unvoiced frames have relatively low frequency in all subbands, corresponding to the spectral energy distribution in a clean speech signal as shown in Fig. 3a. The more corrupt the speech signal is, the more this spectral behavior will change, as seen in Fig. 3b, so the supervised labeling will fix U tags. From this training, the mask estimator will classify as U those regions that are more affected by noise in the test stage.Speaker recognition systems take advantage of that, since most useful speaker information lies in low frequencies, so high frequencies do not contribute

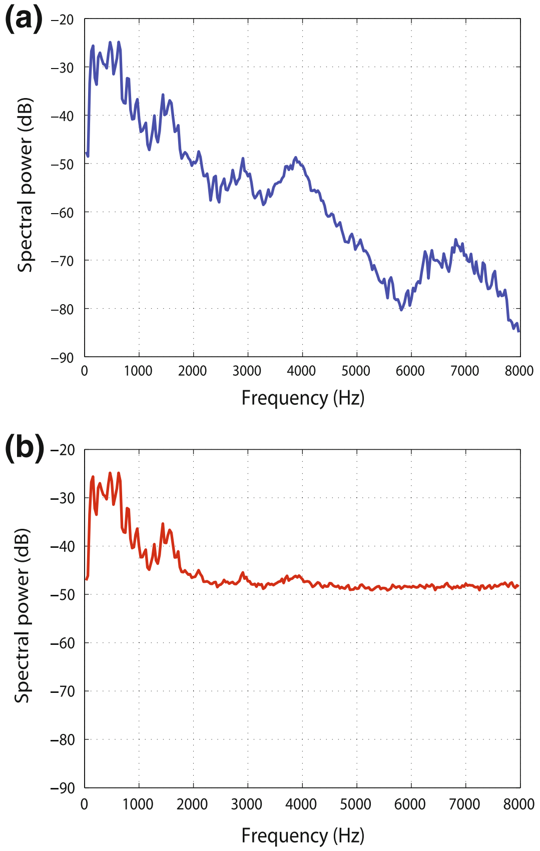


Fig. 3 Spectrum in a single voiced frame of **a** a clean speech signal, **b** a 10 db white noise corrupted speech signal

much in recognizing speakers and at the same time are very sensitive to additive noise corruption.

The second and third features are related to the statistical behavior of signals. Kurtosis is used for capturing information about a signal deviation from gaussianity. Features that represent a clean speech signal generally have a Gaussian distribution, in this case the kurtosis will have a great value. If the clean speech signal is combined with noise its distribution will usually become less Gaussian, in which case its kurtosis would decrease. The mask classifier should therefore select $t\!-\!f$ regions with high kurtosis as more reliable.

The third feature is the energy variance in each $t\!-\!f$ region, with the goal of obtaining flatness. As it is known, the clean speech spectrum is made up of crests and valleys, when speech is corrupted by noise the valleys tend to flatten, as can be seen in Fig. 3. So, the mask classifier is expecting lower variance values for those regions that are corrupted by noise.

Speaker recognition systems are designed assuming the gaussianity of speech features. So the lost of gaussian distribution introduces a variability between the statistical behavior of the test set and the previously trained set, which works against the system's good performance, and causes mistakes in score computation. Hence, were labeled as U those $t\!-\!f$ regions which have lost their gaussianity.

The fourth feature is the likelihood that a specific spectral region has been corrupted by noise. To obtain this, we computed the ratio between the subband energy and the signal noise energy, obtaining noise energy by noise estimation [3]. This technique has the drawback that in the presence of highly non-stationary noise, the signal noise energy estimate will not necessarily be accurate, so it would lack accuracy for some noises. Nevertheless, we decided to use it because it provides a measure of the level of noise that is corrupting the speech signal, and it is a fact that a spectral region could be able to keep the speaker discriminative information depending on the level of noise corruption, among other things. So this is a useful hint for determining spectral reliability in speaker recognition systems. In the case of noise estimation errors, the other mask features must ensure that classifier performance is not affected.

Following this idea, we decided to include as a fifth feature the local SNR too, i.e., the feature normally used in SNR based mask estimation, which is computed similar to what was done in SNR criterion methods proposed by Drygajlo and El-Maliki [21].

The last two features are only computed for voiced speech frames because both are pitch dependent. Due to this precise fact, those are the most striking features for the speaker recognition task. Previous works [27] show that most speaker discriminative information lies on voiced frames, so these mask features contribute to characterizing these frames. The sixth feature is the periodicity and the seventh the relationship between energies at the pitch harmonics and outside the pitch harmonics.

3 Experimental setup

3.1 Corpus

This article evaluates the performance of the FC criterion in mask estimation for the MF approach through a speaker verification experiment, conducted with a set of 100 male speakers from AHUMADA [28], a Spanish NIST 2001 speech database for speaker characterization and identification. To perform the evaluation, the speaker verification system was trained and tested with clean speech to establish the clean baseline; then, for setting the dirty baseline, it was tested with corrupted speech without using the MF approach. Later on, the system was tested with the same corrupted speech used in the dirty baseline but using the MF approach. All speech material used was taken from 3 different Ahumada microphonic sections: M2, M3 for training and M1 for testing. Each of these utterances contains about 90 s of spontaneous speech, making the experiment text independent. All speech material used for training and testing is digitized at 16 bits, at 16,000 Hz sample rate.

The corruption signal comes from four different noise environments:

- stationary white noise
- pseudostationary street noise, which is a mixture of different noises
- music from Guns and Roses band, highly harmonic and non-stationary noise
- babble noise, special case of non-stationary noise, highly correlated with voice because is the voice of other speakers

All those types of noise were added electronically to test speech signals at different SNR levels, from 0 to 20 dB in 5 dB steps.

3.2 Missing feature protocol

The MF approach is divided into 2 steps: missing feature detection and missing feature compensation. For compensation, the classical marginalization technique [29] was used, even taking into account its limitation compared with the use of optimal MFCC features. However, it should be noted that in future any refinements to this method (MF bounds or MF imputation) could be applied to the MF mask estimation method, allowing the use of MFCC and even improving the system performance. We, therefore, consider that using Mel spectral features, rather than Mel cepstrum features is sufficient for the purpose of evaluating the accuracy of the proposed MF mask estimation method. For detection, three types of mask were used:

- a) Oracle masks, to determine the ideal performance that speaker verification could reach using the MF approach.
- b) Spectral Subtraction mask (SS-mask) [21], based on the SNR criterion that allows us to establish a comparative line.
- c) Feature Classification mask (FC-mask), based on the FC criterion, which is the proposal of this article.

To estimate the oracle mask (a), local SNR for each $t{-}f$ region was computed, with a priori knowledge of the noise spectrum, then a threshold of SNR = 0 dB was established, selecting as U the regions whose spectral speech power is inferior to spectral noise power.

SS-mask (b) [21] uses a frame by frame spectral subtraction method as spectral reliability detector based on an estimated noise spectrum. The reliability decision of spectral regions then uses the following rule:

$$\begin{array}{lll} |Y(fr,s)|^2 \leq |\hat{N}(fr,s)|^2 & then & Y(f,s) \leftarrow U \\ |Y(fr,s)|^2 > |\hat{N}(fr,s)|^2 & then & Y(f,s) \leftarrow R \end{array} \quad (1)$$

where Y is the noisy signal and $\hat{N}$ is the noise estimated for each $t{-}f$ region represented by frame (*fr*) and subband (*s*). The Spectral Subtraction algorithm [2] and Martin's noise estimator [4] were used to estimate the noise spectral power. The threshold selected to determine reliability is the same as used above (SNR = 0 dB).

To estimate FC-mask (c), a set of 25,179 speech signals was used for training the Bayesian classifier. These signals were a random selection of short read phrases from 25 speakers of AHUMADA from four different microphonic sections [28], corrupted by stationary pink noise, pseudo-stationary exhibition noise, music noise (theme of Pulp fiction film), and non-stationary restaurant, at SNR levels from 5 to 25 dB, so around 27 h of speech were obtained. As a result, classifiers were trained and later used to estimate the masks used during testing.

3.3 Speaker verification protocol

For applying the MF approach, speech signals were represented with Log-Mel Spectral features: a Hamming window with 25 ms window length and 15 ms of overlap is applied to each frame and a short time spectrum is obtained applying a FFT. Then, 20 Mel filterbanks were applied over it followed by a logarithmic transformation. For implementing the dirty baseline (BSLN-CSp), state of the art MFCC features were used, which were computed according to the process previously described adding the transformation to cepstrum domain and finally selecting 15 cesptral coefficients as features.

A speech set from 50 male speakers from Ahumada's M3 section was used to create a gender dependent Universal Background Model (UBM) [24] using a Gaussian Mixture Model (GMM) of 512 gaussians. The number of mixtures in the GMM was chosen taking into account the number of speakers, the phonetic richness and the signals duration to create the UBM. Other 50 different male speakers were used as targets and their models were obtained by adaptation from the UBM using the Maximum a Posteriori (MAP) approach [30]. The targets speech set was taken from Ahumada's M2 section. For testing were taken 50 speech signals from Ahumada's M1 section corresponding of each of the 50 target speakers. The text contained in speech signals from M1, M2, and M3 sections is different for all speakers, so the speaker verification experiment is text independent. The testing speech set was corrupted by different types of additive noise (Sect. 3.1). The following experiments were conducted:

1. BSLN-Cln/CSp, MF-Oracle: Three speaker verification baselines were trained with clean speech and tested with different speech sets. Clean baseline (BSLN-Cln) was tested with the same clean signals; dirty baseline (BSLN-CSp) was tested using the set of corrupted speech signals specified in Sect. 3.1 and MF with oracle mask baseline (MF-Oracle) was tested applying MF approach with oracle mask and marginalization.
2. MF-SNR: Speaker verification applying MF approach with SS-mask, was trained on clean data and tested with the set of corrupted speech signals specified in Sect. 3.1.
3. MF-FC (Proposed mask criterion): Speaker verification applying the MF approach with the proposed FC-mask, was trained on clean data and tested with the set of corrupted speech signals specified in Sect. 3.1.

All in all, 2,500 trials were done—50 client speakers against each of 50 target models—for each type of noise (white, street, music, babble), SNR level (0, 5, 10, 15, 20 dB) and noise compensation method (without any: MFCC baseline, MF-Oracle, MF-SNR, MF-FC). As a whole, 200,000 trials in 80 experiments were done.

4 Results and discussion

4.1 Mask estimation accuracy in regard to oracle mask

Figure 4 presents a comparison of speaker verification experiment results in EER[3] percentage, tested using speech

[3] EER: The error rate of a verification system when the operating threshold for the accept/reject decision is adjusted such that the probability

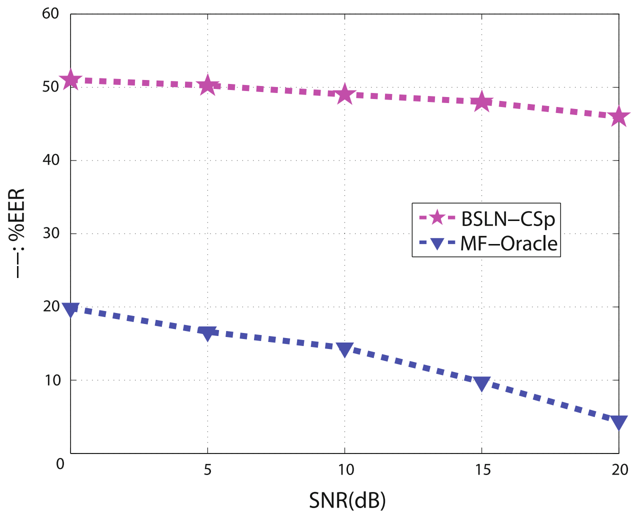


Fig. 4 Speaker verification with corrupted speech (BSLN-CSp) and applying missing feature with oracle mask (MF-Oracle)

Table 1 Average of hit indexes of SS and FC-masks computed in regards to oracle mask, obtained in percentage for each corruption category: speech corrupted by white, street, music and babble noise, at 0, 5, 10, 15, 20 dB of SNR levels

SNR (dB)	Hit: white		Hit: music		Hit: street		Hit: babble	
	SS	FC	SS	FC	SS	FC	SS	FC
20	58	86	57	76	72	73	69	76
15	55	88	48	80	67	76	61	78
10	54	90	39	83	62	79	51	78
5	56	92	32	85	56	81	42	79
0	57	92	27	86	50	82	35	82

corrupted by white noise at different SNR levels applying MF approach with oracle mask (MF-Oracle) and without it (BSLN-CSp). It shows that the EER percentage decreased a lot in MF-oracle mask application, consistent with oracle mask definition as an ideal mask.

These results encourage us to use oracle mask as a comparative pattern for other mask performance, computing the amount of tagged t–f regions that match with the oracle mask, which we called hit index, composed by the matched R and U regions. Both masks were computed with all sets of corrupted speech and the results of each mask for each corruption category were ranked, showing a summary of results in Table 1.

In Table 1 the stability in hit indexes for each type of noise obtained from the FC-mask can be appreciated, while for the SS-mask the hit index values tend to improve with the improvement of SNR, which denotes a strong dependency of SS-mask accuracy with the SNR level. This fact leads us to conclude that FC-mask is more robust to noise than SS-mask.

In general, Table 1 shows that hit percentage of the proposed FC-mask is consistently greater than for the SS-mask. These results suggest that in general the proposed FC-mask will outperform SS-mask in speaker verification experiments. These results could give us a preview of the performance of a given spectrographic mask estimator without the need to carry out any recognition experiment. This hit method for comparing different masks against the oracle mask, could therefore be very useful in MF techniques development. This hypothesis will be analyzed in the following section by comparison of hit indexes and speaker verification results.

4.2 Speaker verification results

Figure 5 shows results of speaker verification experiments, as described in Sect. 3.3, in EER percentage versus SNR, using speech corrupted by four types of noise at five SNR levels, applying the MF approach with an oracle mask (MF-Oracle), the SNR criterion's mask (MF-SNR) and the proposed mask criterion (MF-FC). Some general conclusions could be obtained from the results:

- When SNR increases, mainly SNR > 15 dB, the usefulness of any MF approach for speaker verification decreases, although in general it still outperforms speaker verification results under corrupted speech without any mask. This happens because if the power of noise is low, EER results tend to those values that could be obtained if the speaker verification had been carried out with clean speech. This is a very common behavior for noise compensation methods applied to high SNR speech in speaker verification, that could be seen in Raj and stern [12] and El-Maliki and Drygajlo [15] too.
- On the other hand, generally MF-FC reaches better performance than MF-SNR—lower EER for the same SNR—with the exception of babble and street noise at SNR > 15 dB and music noise at SNR = 20 dB, that will be analyzed later. The results of MF-FC are due to the fact that it takes into account many features to determine spectrum reliability, unlike MF-SNR. In spite of that, the performance is different for each type of noise, given the very different kinds of noise used.

Some noise-related conclusions could be obtained:

- When a noise compensation method is used with the recognition system, generally is easier to compensate for the effects of stationary noise than for non-stationary [30]. In speaker verification experiment under white noise

Footnote 3 continued
of false acceptance and that of false rejection become equal. This measure is commonly used as recognition accuracy measure in biometrics applications such as speaker verification.

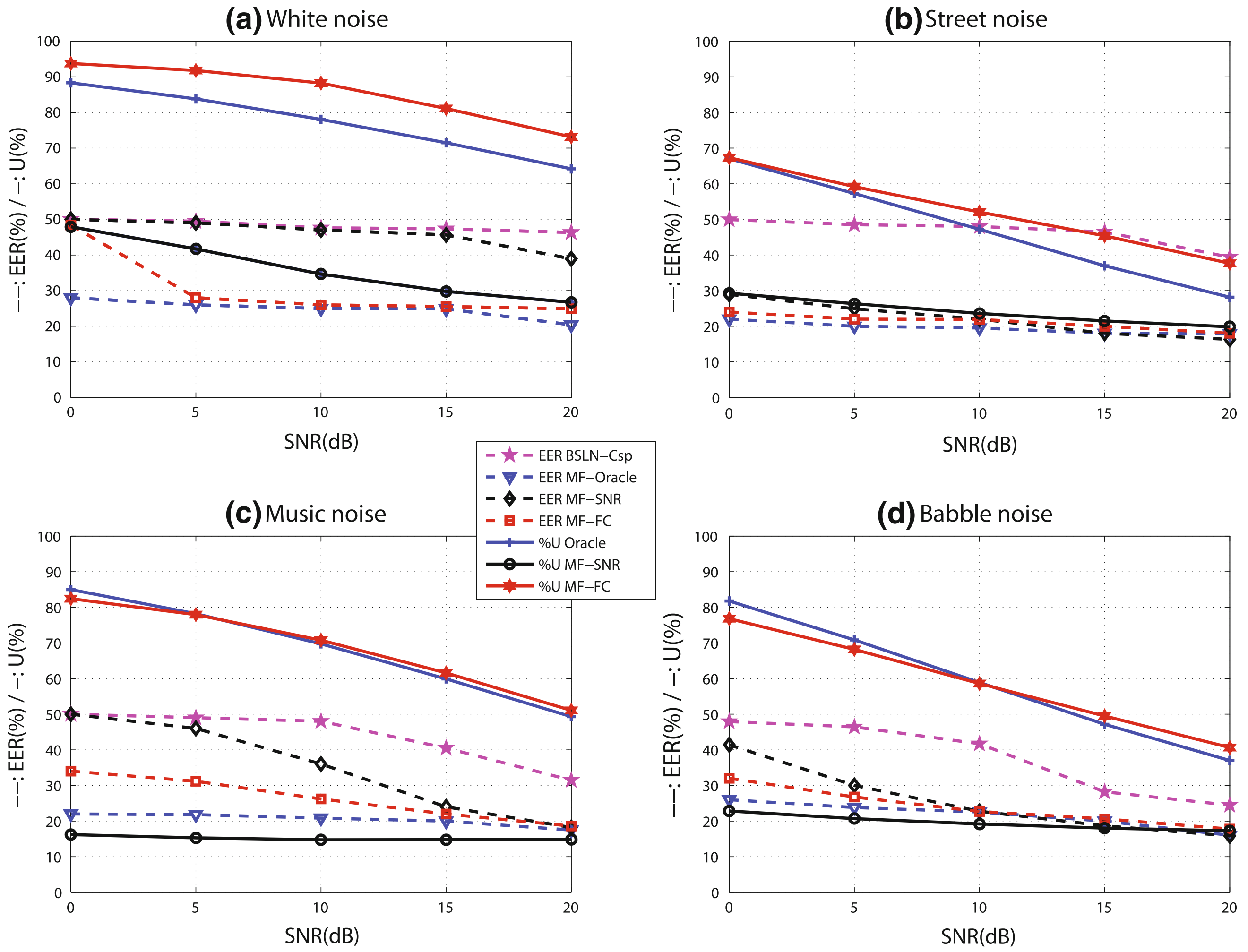


Fig. 5 EER percentage of speaker verification experiments and *U* regions percentage versus different SNR with speech corrupted signals by **a** white, **b** street, **c** music and **d**) babble noise

(Fig. 5a) MF-FC behavior is very close to MF-Oracle behavior, because it is stationary noise. On the other hand, this is an expected result because of the high hit percentage of FC-mask, for all SNR levels, shown in Table 1.

- In speaker verification experiment under street noise (Fig. 5b), in general MF-SNR improves EER compared to BSLN-CSp, but MF-FC performs better. In comparison with MF-SNR, MF-FC improves EER for the most difficult SNR conditions, however for SNR > 15 dB MF-SNR performs a little better. This happens since the Martin's noise estimation method used is limited when dealing with a pseudo-stationary noise, and the influence of the behavior of noise compensation methods for corrupted signals with a low level of noise corruption explained above.
- In speaker verification experiment under music noise (Fig. 5c), MF-SNR and MF-FC had a similar behavior as in street noise.
- In speaker verification experiment under babble noise (Fig. 5d), for the most challenging SNR levels (0, 5, 10 dB) MF-FC outperforms MF-SNR. In this special case of non-stationary babble noise, MF-FC must differentiate between impostor and target voices, because babble noise is one of the most challenging noise interferences for all speech systems [31]. This result shows us that spectral features selected to estimate theFC-mask are not speaker discriminative enough to handle babble noise.
- There is no complete correspondence between hit index and EER results due to the differences at SNR = 15, 20 dB of street noise; SNR = 15, 20 dB of babble and street noise and SNR = 20 dB of music noise. The mismatched results are due to the fact that hit index analysis was too rough for speaker recognition applications, taking into account only the quantity of hits, without evaluating the spectral region of each hit. In fact, a hit in a spectral region with speaker discriminative information impacts more in EER than a hit in spectral region without any speaker discriminative information.

Table 2 Speaker verification results expressed in EER percentage for the proposed MF-FC and MF-SNR, oracle mask and baselines

Noises	SNR (dB)	BSLN-CSp	MF-oracle	MF-SNR	MF-FC
BSLN-Cln	30–40	24	–	14	22
White	20	46.32	20.36	38.93	24
	15	47.34	24.85	45.63	25
	10	47.59	24.98	47	26
	5	49.46	26	49	28
	0	50	28	50	48.4
Street	20	39.38	17.91	16.28	18
	15	46.48	18	18	20
	10	48	19.5	22	21.91
	5	48.5	20	24.89	22
	0	49.92	22	28.93	24
Music	20	31.46	17.46	18.2	18.61
	15	40.48	20	24	22
	10	48	20.81	36	26.2
	5	49	21.79	46	31.18
	0	50	22	50	34
Babble	20	24.48	16	15.83	17.75
	15	28.2	20	18.65	20.61
	10	41.71	22.53	22.73	22.69
	5	46.44	23.79	30	26.77
	0	47.91	26	41.38	32

Table 2 presents a summary of speaker verification effectiveness in EER values, obtained in the experiments. As a reference the first line shows results for training and testing with clean speech. This table shows that MF-FC mask offers the best speaker verification results, under highly contaminated noise conditions (SNR <10 dB), for all type of noises.

4.3 Speaker verification accuracy (EER) and amount of *U* regions versus SNR

To conclude the discussion, we analyzed the relation between speaker verification accuracy (through EER) and the proportion of *U* regions for each mask and each type and level of noise. So, the percentage of *U* regions was computed for each mask corresponding to each type and level of noise. Figure 5 presents the relation between speaker verification accuracy, SNR level and percentage of *U* regions, for the three evaluated mask estimation criteria.

Results for the oracle mask are the reference for the others, since this mask shows the lowest EER results for any type and level of noise. It is quite clear that the relation between the proportion of *U* regions and EER behavior is particular to for each type of noise, however for all of them the amount of *U* regions increases with the SNR decrease, as expected. It is noticeable that the SS-mask detects the smaller amount of *U* regions between all masks and in most cases reaches the worst EER. This fact supports the MF hypothesis that it is better for speaker verification performance, to process incomplete noisy spectrum than the whole noisy spectrum.

On the other hand, oracle mask tends to be inversely proportional regarding SNR level, with an important slope. However the SS-masks have lower slope, tending to have a constant behavior instead. This happens because the noise estimation method does not deal properly with non-stationary noise, which influences the reliability decision. The SS-mask uses this as the only measure to take into account when tagging a region as *U* or *R*. Figure 5 also shows that the FC-mask's *U* percentage curves maintain similar behavior to oracle curves, detecting many more *U* regions than SS-mask, more than 30 %. Those facts demonstrate that the FC criterion outperforms the SNR criterion in determining spectrum's reliability of speech. This explains why the FC criterion provides more accurate speaker recognition performance under noisy conditions.

On the other hand, EER results obtained applying MF-oracle mask at SNR = 20 dB to all types of noises are very low, which denotes high accuracy in speaker verification. The amount of *U* regions obtained at this point indicates that there is a minimum value of *U* regions, for which the verification results are good. Thus, there will be a part of the spectrum (the minimum percentage of *U* regions) that contributes to successful speaker verification, which means that it does not have speaker discriminative information. This could be very useful to improve the speed of the speaker verification process, analyzing just a reduced number of spectro-temporal regions.

5 Conclusions and future work

This article is aimed at dealing with the problem of robust speaker recognition in speech corrupted on different noise environments. We proposed the use of a Feature Classification criterion to estimate MF masks in speaker recognition, using a method of Seltzer et al. [23] previously used for speech recognition. For that proposal we were based on the fact that the SNR criterion—which has been the most frequently used to estimate masks in this task—does not yield the most effective results. We believe that this problem is due to the fact that the SNR criterion is only based on SNR estimation, which has the goal of enhancing speech signal, and by doing this, it could remove speaker-distinguishing information. On the other hand, FC criterion make use of wider context spectral information, which is more advantageous to face noise than use only SNR.

To conclude, the evaluated mask estimation criterion—MF-FC—has the advantage of using different spectro-temporal measures not dependent on one another. So, if any part of this information is affected by noise, the rest could ensure that reliability decision remain consistent. On the other hand, MF-FC's computational cost is higher than MF-SNR. This could possibly reduced using $t-f$ patches corresponding to subbands with speaker discriminative information [32,33] instead of all $t-f$ regions.

We explored a mask estimator quality measure: the hit index, presupposing that this could give a preview of the performance of a given spectrographic mask estimator without the need to carry out any recognition experiment. However, in spite of the fact that most results matched with EER behavior, they did not correlate perfectly with EER. Hence, we think that the problem is that the analysis related to hit index only took into account the quantity of hits. However in speaker recognition applications a hit in a spectral region with speaker discriminative information impacts more in EER than a hit in spectral region without any speaker discriminative information. So, in the future, we will work on a mask estimator quality measure based on the computation of hit index, but taking into account this idea.

We evaluated our proposal through speaker verification experiments. The experiments demonstrated that, for speech corrupted by stationary, pseudo-stationary and some non-stationary noises, MF-FC outperforms the MF-SNR mainly when speech corruption increases (SNR < 10 dB).

Finally, we analyzed the relation between speaker verification accuracy and the proportion of U regions. From this, we concluded that there is a part of the spectrum that contributes a little to successful speaker verification. This could be very useful to make a computing reduction of speaker verification process by analyzing just a reduced amount of speaker feature vectors. This would depend on the spectral distribution of the signal, and some spectral parameters could be used as reference of spectral regions with potentially useful speaker information, such as pitch and formants.

The analytical conclusions and experimental results obtained in this article, encourage us to continue using MF-FC. As future work, we could use other features more related to the speaker identity, with the idea of associate the reliability decision with the corruption of $t-f$ regions which have useful speaker recognition information.

References

1. Benesty, J., Sondhi, M.M., Huang, Y.: Springer Handbook of Speech Processing. Springer, Berlin (2008)
2. Berouti, M., Schwartz, R., Makhoul, J.: Enhancement of speech corrupted by acoustic noise. In: IEEE ICASSP (1979)
3. Hirsch, H.G., Ehrlicher, C.: Noise estimation techniques for robust speech recognition. In: ICASSP (1995)
4. Martin, R.: Noise power spectral density estimation based on optimal smoothing and minimum statistics. In: IEEE Transaction on Speech and Audio Proceedings, vol. 9 (2001)
5. Teunen, R., Shahshahani, B., Heck, L.P.: A Model-Based Transformational Approach to Robust Speaker Recognition. ICSLP, Beijing (2000)
6. Davis, S., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In: IEEE Transaction Ac. Speech, and Signal Processing, vol. 28, issue number 4, pp. 357–366 (1980)
7. Hermansky, H.: Perceptual linear prediction (PLP) analysis for speech. J. Acoust. Soc. Am. **87**(4), 1738–1752 (1990)
8. Gales, M.J.F., Young, S.J.: HMM recognition in noise using parallel model combination. In: EUROSPEECH'93, pp. 837–840 (1993)
9. Sagayama, S., Yamaguchi, Y., Takahashi, S., Takahashi, J.: Jacobian approach to fast acoustic model adaptation. In: ICASSP (1997)
10. Bimbot, F., Bonastre, J.-F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-Garcia, J., Petrovska-Delacretaz, D., Reynolds, D.A.: A tutorial on text-independent speaker verification. Eurasip J. Appl. Signal Process. **4**, 430–451 (2004)
11. Kinnunen, T., Li, H.: An overiew of text-independent speaker recognition: from features to supervectors. Speech Commun. **52**, 12–40 (2010)
12. Raj, B., Stern, R.: Missing-feature approaches in speech recognition. In: IEEE Signal Processing Magazine (2005)
13. Raj, B., Seltzer, M., Stern, R.M.: Reconstruction of MFs for robust speech recognition. Speech Commun. **43**, 275–296 (2004)
14. El-Maliki, M., Drygajlo, A.: Integration and imputation methods for unreliable feature compensation in GMM based speaker verification. In: Speaker Recognition Workshop Odyssey, Crete, Greece (2001)
15. El-Maliki, M., Drygajlo, A.: Missing Features Detection and Handling for Robust Speaker Verification. Eurospeech, Budapest (1999)
16. Demange, S., Cerisara, C., Haton, J.-P.: Accurate Marginalization Range for Missing Data Recognition in Interspeech. Interspeech, Antwerp (2007)
17. Padilla, M., Quatieri, T., Reynolds, D.: MF Theory with Soft Spectral Subtraction for Speaker Verification. Interspeech, Pittsburgh (2006)
18. Ming, J., Hazen, T., Glass, J.R., Reynolds, D.A.: Robust speaker recognition in noisy conditions. IEEE Trans. Speech Audio Process. **15**, 1711–1723 (2007)
19. Pullella, D., Kuhne, M., Togneri, R.: Robust speaker identification using combined feature selection and missing data recognition. In: ICASSP (2008)
20. Cerisara, C., Demange, S., Haton, J.-P.: On noise masking for automatic missing data speech recognition: a survey and discussion. Comput Speech Lang **21**(3), 443–457 (2007)
21. Drygajlo, A., El-Maliki, M.: Speaker verification in noisy enviroments with combined spectral subtraction and MF theory. In: Signal Processing Laboratory, Swiss Federal Institute of Technology at Lausanne (1998)
22. Shao, Y., Wang, D.: Robust speaker recognition using binary time-frequency masks. In: ICASSP (2006)
23. Seltzer, M., Raj, B., Stern, R.M.: A Bayesian classifier for spectrographic mask estimation for MF speech recognition. Speech Commun. **43**, 379–393 (2004)
24. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker verification using adapted gaussian mixture models. Digit Signal Process **10**, 19–41 (2000)
25. Talkin, D.: "A Robust Algorithm for Pitch Tracking (RAPT)", Speech Coding and Synthesis. Elsevier, Amsterdam (1995)
26. Duin, R.P.W., Juszczak, P., Paclik, P., Pekalska, E., de Ridder, D., Tax, D.M.J., Verzakov, S.: "PRTools4 A Matlab Toolbox for Pattern Recognition", Version 4.1, Delft Pattern Recognition Research Faculty EWI—ICT, http://prtools.org/ (2007)

27. Zilca, R., Kingsbury, B., Navratil, J., Ramaswamy, G.: Pseudo Pitch Synchronous Analysis of Speech with Applications to Speaker Recognition. In: IEEE Trans. Audio Speech Lang. Process. **14**, 467–478 (2006)
28. Ortega, J., Gonzalez, J., Marrero, V.: AHUMADA: A Large Speech Corpus in Spanish for Speaker Characterization and Identification. Speech Commun. **31**, 255–264 (2000)
29. Drygajlo, A., El-Maliki, M.: Speaker Verification in Missing Features Detection and Handling for Robust Speaker Verification. EUROSPEECH, Budapest (1999)
30. Davis, G.M.: Noise Reduction in Speech Applications. CRC PRESS LLC, New York (2002)
31. Krishnamurthy, N., Hansen, J.H.L.: Babble noise: modeling, analysis, and applications. In: IEEE Trans. Audio Speech Lang. Process. **17**(7), 1394–1407 (2009)
32. Besacier, L., Bonastre, J.-F.: Subband architecture for automatic speaker recognition. Signal Process. **80**, 1245–1259 (2000)
33. Besacier, L., Bonastre, J.F., Fredouille, C.: Localization and selection of speaker-specific information with statistical modeling. Speech Commun. **31**, 89–106 (2000)
34. Morris, A.C., Green, P.M.: Some solutions to the missing feature problem in data classification with application to noise robust ASR. In: ICASSP, pp. 737–740 (1998)