RIELAC
Revista de Ingeniería Electrónica,
Automática y Comunicaciones

New Missing Features Mask Estimation Method for Speaker Recognition in Noisy Environments

Dayana Ribas González and José R. Calvo de Lara

ABSTRACT/RESUMEN

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 Missing Feature (MF) approach to compensate for noise. In most of those applications the spectral reliability decision step is done using the Signal to Noise Ratio (SNR) criterion. This has the goal of enhancing signal power rather than noise power, which could be dangerous in speaker recognition tasks, because useful speaker information could be removed. This work proposes a new mask estimation method based on Speaker Discriminative Information (SDI) for determining spectral reliability in speaker recognition applications based on the MF approach. The proposal was evaluated through speaker verification experiments in speech corrupted by additive noise. Experiments demonstrated that this new criterion has a promising performance in speaker verification tasks.

Key words: Formants, Missing features approach, Speaker recognition.

En la actualidad, muchas aplicaciones de reconocimiento de locutores deben manejar voz corrupta por ruido aditivo ambiental sin tener conocimiento previo sobre las características del ruido. Trabajos previos de reconocimiento de locutores han usado la teoría de Rasgos Perdidos (MF: Missing Features) para compensar el ruido. En muchas de estas aplicaciones el paso de la decisión de confiabilidad espectral se hace usando el criterio de Relación Señal a Ruido (SNR: Signal to Noise Ratio). Este tiene el objetivo de resaltar la potencia de señal sobre la potencia de ruido, lo que pudiera ser peligroso en tareas de reconocimiento de locutores, porque se pudiera eliminar información útil del locutor. Este trabajo propone un nuevo método de estimación de máscara basado en Información Discriminativa del Locutor (SDI: Speaker Discriminative Information) para determinar la confiabilidad espectral en aplicaciones de reconocimiento de locutores basadas en la teoría de MF. La propuesta fue evaluada en experimentos de verificación de locutores con voces corruptas por ruido aditivo. Los experimentos demostraron que este criterio tiene un desempeño prometedor en verificación de locutores.

Palabras Claves: Formantes, Teoría de rasgos perdidos, Reconocimiento de locutores

Nuevo método de estimación de máscara de rasgos perdidos para reconocimiento de locutores en ambientes ruidosos

INTRODUCTION

Plenty of advances have been made in automatic speaker recognition (ASR) technology in the last decades. Despite that, robust speaker recognition in noisy environments is even a challenging task for the current technology [1].

Robust ASR in noisy environments ought to have great attention for the researchers because is a very common case in applications. For example, in forensics there is a current trend to implement auditory and semiautomatic analysis over telephone conversations for recognizing persons in a conversation [2]. In addition is the speaker diarization, which consists in determining who spoke in each moment, is a special case of speaker recognition in noisy environments, where the voice of other persons is the acoustic noise, this kind of noise is called babble noise and is very difficult to deal with [3]. In remote access services, identity verification using user's voice is advantageous for both, users and providers, because of the security that could offer a biometric measure in regard of text password or pin, which could be stolen or cracked easily. What is more, the operation and personalization of those services could be done speaking. This requires the integration of speech and language recognition technologies besides speaker recognition, but will provide a great range of automatization for those services. There are several other applications of ASR technology, however these examples are enough to demonstrate the importance of working on strengthening the speaker recognition scheme when dealing with voices acquired in noise environments.

Missing feature approach [4] has been applied to robust speaker recognition in noisy environments to compensate for noise, with promising results [5-7]. This approach is based on the fact that any noise affects time-frequency (t-f) components of speech spectrum in different ways, so it consists in detecting spectrum corruption level and determining which part of the spectrum is reliable to be used in recognition.

The use of MF approach in speaker recognition has two steps. Firstly the detection of the reliability degree of corrupted speech spectrum, by creating a map of reliability in correspondence with t-f components, called spectrographic mask. The mask is formed by reliable (R) and unreliable (U) labels that correspond to each t-f component in the spectrum, so the analysis includes each t-f component in the spectrum. Components which are highly corrupted by noise are tagged with U labels and components with a low level of corruption with R labels. Secondly the missing feature compensation which is based on spectrographic mask. This procedure has two options: to reconstruct unreliable components to perform recognition with the newly reconstructed spectrum or to bypass unreliable components, so as not to use it in the recognition process. The first option submits unreliable reconstruction techniques, called imputation techniques, developed for speech recognition, 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.

As it could be seen, the potential for improvement increases mainly depending on mask estimation accuracy. This happens because missing feature compensation works only with U components determined by mask, if the mask is not accurate, the error is dragged, i.e. some R components will be compensated, while some U components will be kept untouched. In short, it could be said that mask estimation is the most important process in the MF approach, so in this paper we will focus on the mask estimation step.

The paradigm that has been behind most spectrographic masks estimation methods used in speaker recognition consists on determining if a t-f component is dominated by speech or by noise. This is achieved by SNR computation, wherein noise energy estimation for each t-f component is compared with an estimation of clean signal energy through local SNR computation, and then a threshold is used to determine the spectrum reliability. This paradigm that we will call “SNR criterion” is the basis of most mask estimation methods used in speaker recognition works, where the key is the way to obtain the elements to compute local SNR. An example is the highly used method proposed by Maliki and Drygajlo in 1998 [8], which employs Spectral Substraction; another example is the method proposed in [9] which uses Minimum Mean Square Error (MMSE). Those methods perform accurately for stationary noise, but it degrades severely in non-stationary noise conditions, thus other methods have been developed over the SNR criterion. The work presented in [10] is an example of that, this is specially designed for the highly difficult and non-stationary babble noise, which is based on a speech segregation system using a pitch estimator to discern between target and impostor speech, then are selected the t-f components dominated by target as R components of mask. Another is the Pullella'set. al [7] proposal which reuses [8] method and is tuned using a features selection method based on a multicondition training.

In general, methods based on SNR criterion works quite well, some even for non-stationary noises, but have the limitation that all their performance relies only on one feature, the SNR so, the system’s performance will depend on signal and noise estimations accuracy, for non-stationary noises it is quite difficult to achieve accurate estimations [11]. What is more, SNR is not the main criterion used by ASR for recognizing speakers, those systems are based on Speaker Discriminative Information (SDI) instead, that would be affected by SNR index but it is not precisely the key.

In view of these facts, this paper proposed a new mask estimation method for speaker recognition which copes with the limitations suffered by methods based on SNR criterion. This new method employs a SDI criterion and a feature classification algorithm, wherein the reliability of each t-f component is characterized with several features which are speaker discriminative, so the paradigm of this new mask lies on determining if a t-f component keeps enough SDI to be useful in speaker recognition process, where the SDI of a t-f component could be mainly affected by additive noise corruption.

The key contributions of this paper are firstly the introduction of a new concept in the paradigm of spectrographic estimated masks for MF approach fixed to speaker recognition tasks. Moreover, a new mask estimation method is proposed which support the paradigm exposed, considering the SDI as the main measure in the reliability decision. The proposal is evaluated experimentally showing its promising performance over [8] method based on SNR criterion, which is a very used baseline in previous similar works.

From now on section 2 presents the mask estimation method proposed. Section 3 presents an evaluation of the proposal through a speaker verification experiment. Section 4 shows the results obtained with a discussion of those and finally section 5 carry out some conclusions of the whole study.

PROPOSED METHOD

Hypothesis

Formants estimator methods tend to fail when processing speech signals corrupted by noise. These failures consists in detecting false formants or omitting formants in spectral region where actually there are. Therefore we decided to take advantage of this issue, designing a mask estimation method which used mistakes of formants estimator methods as noisy t-f components detector. On the other side, previous works have demonstrated that formants are a valuable SDI in the process of speaker recognition [12], so if formant information could not be accurately recovered from spectrum, hardly this t-f component will have a favorable impact in posterior speaker recognition process. Thus, we can conclude that determining unreliable components from formants corruption ensures that speaker recognition will do only with spectral region with enough SDI, such that it have been capable to survive acoustic noise effects.

The method

The mask estimation method was designed as a supervised classification schema (fig. 1) with two stages:

1. Creation of formant models
2. Computation of spectrographic masks

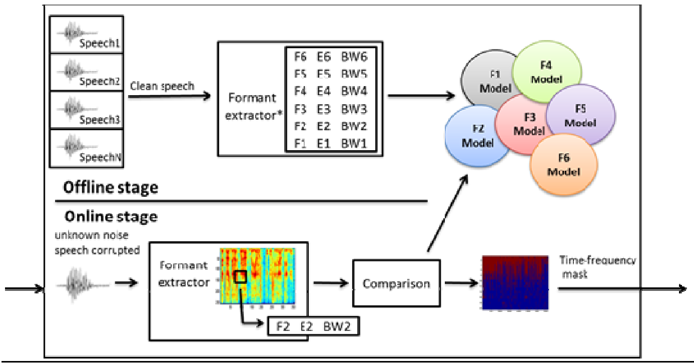

Fig. 1. Schema of mask estimator method proposed

As it is based on formant information was called: FMT mask.

Formant models creation

The first stage is executed offline to the process of creating spectrographic masks, in this are modeled the six formants in separated models. To represent formants were used the frequency (F), energy (E) and bandwidth (BW) of formants as features. They were extracted from frames of several samples of clean speech. To describe the distribution of those features, histograms for all F samples for each formant were calculated, showing the gaussian distribution of formant frequencies, which encourage us to use Gaussian Mixtures Models (GMM) to model the formants. Then histograms of E and BW samples by formant were computed to fix the number of gaussian needed for each specific GMM, staying as indicate table I.

Table I
Number of gmm mixtures for each formant model

Formants	1	2	3	4	5	6
Mixtures	1	2	2	3	1	1

Let *FMT* denote a formant vector, which will be composed by *F, E* and *BW* measure of a specific formant (f) in a frame of speech signal, defined as: $FMT_f = (F_f, E_f, BW_f)$. Given a collection of *FMT*, GMM parameters are estimated using the iterative expectation maximization (EM) algorithm [13]. The EM algorithm iteratively refines the maximum likelihood models parameters to monotonically increase the likelihood of the estimated model for the observed formant vectors. The EM equations for training a GMM can be found in [14]. The formants models are denoted as: λ_f, where f is the number of formant. We can assume formant vectors statistically independents, so the loglikelihood of a model λ_f for a sequence of formant vectors, to specific formant, is computed as:

$$logp(FMT|\lambda) = \sum_{fr=1}^{Frames} log\, p(\overrightarrow{fmt}_{fr}|\lambda) \qquad (1)$$

where:

$$p(\overrightarrow{fmt}_{fr}|\lambda) = \sum_{m=1}^{Mixtures} w_m\, p(\overrightarrow{fmt}_{fr}) \qquad (2)$$

$$p(\overrightarrow{fmt}_{fr}) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma_{fr}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\overrightarrow{fmt}-\vec{\mu}_{fr})'(\Sigma_{fr})^{-1}(\overrightarrow{fmt}-\vec{\mu}_{fr})} \qquad (3)$$

fr: frame
μ_{fr}: distribution's mean
Σ_{fr}: distribution's variance
w_m: distribution's weight

Spectrographic mask computation

The second stage takes over computation of spectrographic masks by determining reliability of t-f components to label as R/U for creating the mask. For that, the FMTf for each formant in each frame of target speech signal are extracted and those are compared with the formant models to obtain the loglikelihood of FMTf regarding all formant models. If the maximum loglikelihood of FMTf belongs to the corresponding λf (for example: FMT2 = λ2) then this t-f component is labeled as R otherwise it is U (eq. 4).

$$\begin{array}{lll} FMT_f \in \lambda_f & \rightarrow & mask(t,s) = R \\ FMT_f \notin \lambda_f & \rightarrow & mask(t,s) = U \end{array} \qquad (4)$$

Formant tracking method

For both,online and offline stages is necessary to extract FMT_f. So we designed a method for obtaining F_f, and measure corresponding $E_f$and BW_f. This method is based on spectral phase acquired from Chirp Group Delay (CGD) function [15].
Firstly CGD of the speech signalis computed, then the frequencies corresponding to the peaks of CGD are evaluated for selecting which one corresponds to each formant and F_f is obtained according eq. 5:

$$Prob_f = \max\,(d_f * cos <) \tag{5}$$

where:

- $Prob_f$: probability of a CGD function peak belonging to the formant *f*

$$d_f = \frac{1}{|freq_{pick} - CFreq_f|} \tag{6}$$

- d_f: Measures how near is the frequency of CGD peak analyzed to central frequency of formant subbands.
- *CFreq*: central frequency of formant subbands (F1 = 500 Hz, F2 = 1500 Hz, F3 = 2750 Hz, F4 = 4000 Hz, F5 = 5000 Hz, F6 = 6000 Hz)
- *freq*: frequency of CGD peak

$$\cos < = \frac{|fr_{pick} - fr_{pick-1}|}{|freq_{pick} - freq_{pick-1}|} \tag{7}$$

- *fr:* frame of CGD peak

The energy is computed for each frequency taking the spectral intensity. The bandwidth is computed taking the frequencies with ± 3 dB of attenuation from spectrogram using corresponding formant frequency as reference, these two values are subtracted for obtaining the bandwidth.

Experimental setup

Corpus

This article evaluates the performance of FMT mask estimator of MF approach through a speaker verification experiment, conducted with a set of 100 male speakers of AHUMADA [16], 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 dirty baseline but using the MF approach. Each train and test utterances contains about 90 seconds of spontaneous speech from the Ahumada'smicrophonic section M1. All speech material used for training and testing is digitized at 16 bits, at 16000 Hz sample rate.
The corruption signal comes from a special case of non-stationary noise, called babble noise, which is highly correlated with voice because it is the voice of other speakers. This was added electronically to test speech signals at different SNR levels, from 0 to 20 dB in 5 dB step.

Missing feature protocol

The MF approach is divided into 2 steps: missing feature detection and missing feature compensation. For unreliable compensation, classical marginalization technique [17] was used. This simple method was selected considering our intention of focus on mask estimator performance and taking into account the good results reached by this method in previous works [5]. For detection three types of masks were used:

a) Oracle masks , to determine the ideal performance that speaker verification could reach using MF approach.

b) Spectral Subtraction mask (SS-mask), based on SNR criterion that allows us to establish a comparative line.

c) FMT mask (FMT-mask), based on SDI criterion, which is the proposal of this article.

To estimate FMT-mask (c), a set of clean speech signals was first selected to create the formant models. The signals were short read phrases from 50 speakers of AHUMADA [16] from 3 microphonic and 3 telephonic sections, in total 300 signals, around 20 minutes of speech for each model.

Speaker verification protocol

For applying MF approach, speech signals were represented with Log-Mel Spectral features: a Hamming window with 20ms window length and 10ms of overlap is applied to each frame and a short time spectrum is obtained applying a FFT. Then 20 Mel filterbank were applied over it followed by a logarithmic transformation. For implementing "dirty baseline" state of the art MFCC features were used, computed according to the process described previously, adding the transformation to cepstrum domain and finally selecting 15 cesptral coefficients as features.

Speech from 50 male speakers were used to create a gender dependent Universal Background Model (UBM) [18] using a GMM of 512 gaussians. The amount of mixtures to GMM was chosen taking into account the number of speakers, the phonetic richness and the signals duration to create the UBM. Other 50 male speakers were used as targets and their models were obtained adapting UBM with Maximum a Posteriori (MAP) approach. Based on our goal the mismatch between train and test sessions produced only by additive noise was measured. We did not introduce any other mismatch source, hence was used for test the same signals used for train in a text-dependent speaker verification, with the difference that those were clean in train session and corrupted by babble noise in test session.

All in all, 2500 trials - 50 client speakers against each of 50 target models – were done for each SNR level (0, 5, 10, 15, 20 dB) and noise compensation method (without any: MFCC baseline, MF-Oracle, MF-SNR, MF-FMT). In total 50000 trials in 20 experiments were done.

RESULTS AND DISCUSSION

Table II presents a summary of speaker verification effectiveness in EER percentage vs. SNR reached by the experiments described in the previous epigraph

Table II
Speaker verification results expressed in EER percentage.

SNR/EER	MFCC	MF-Oracle	MF-SNR	MF-FMT
20	3.22	2.44	3.22	5.55
15	5.10	4.16	4.97	6.04
10	7.67	6	7.22	7.509
5	25.79	6.44	20	14.53
0	44	10	42.73	34.04

The table shows that MF-FMT mask offers the best speaker verification results, under highly contaminated noise conditions (SNR<10dB), however when SNR increases, MF-FMT results are not better than MF-SNR results. 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 [4, 9] too. On the other hand, those results show that only formants are not capable of providing enough SDI to reach MF-Oracle performance, so adding other features with SDI could improve the performance.

CONCLUSIONS AND FUTURE WORK

In spite of babble noise it is very difficult to handle for compensation methods, due to its non-stationarity and that it is very correlated with voice, the proposed mask estimation criterion -MF-SDI- outperformed the MF-SNR in the most difficult conditions, SNR<10 dB.

The analytical conclusions and experimental results obtained in this article encourage us to continue using SDI criterion to create mask estimation methods, as long as we explore other features more related with the speaker identity, to associate the reliability decision with the measure of corruption in information useful for speaker recognition. Futureworkwill be in thisdirection. .

REFERENCES

1.Kinnunen, T. and H. Li, *An overiew of text-independent speaker recognition: From features to supervectors.* Speech communication, 2010. **52**: p. 12-40.
2.Campbell, J.P., et al., *Forensic Speaker Recognition*, in *IEEE Signal Processing Magazine*. 2009. p. 95-103.
3.Tranter, S. and D. Reynolds, *An Overview of Automatic Speaker Diarisation Systems.* IEEE Transactions on Speech and Audio Processing, 2006. **14**: p. 1557-1565.
4.Raj, B. and R.M. Stern, *Missing-Feature Approaches in Speech Recognition*, in *IEEE Signal Processing Magazine*. 2005. p. 101-116.
5.Padilla, M.T., T.F. Quatieri, and D.A. Reynolds, *Missing Feature Theory with Soft Spectral Subtraction for Speaker Verification*, in *Interspeech* 2006, : Pittsburgh, Pennsylvania.
6.Ming, J., et al., *Robust Speaker Recognition in Noisy Conditions.* IEEE Transactions on Audio, Speech and Language Processing 2007. **15**: p. 1711-1723.
7.Pullella, D., M. Kuhne, and R. Togneri. *Robust Speaker Identification Using Combined Feature Selection and Missing Data Recognition*. in *International Conference in Acoustics, Speech and Signal Processing (ICASSP-08)*. 2008. Las Vegas, NV.
8.Drygajlo, A. and M. El-Maliki, *Speaker Verification in Noisy Enviroments with Combined Spectral Subtraction and Missing Feature Theory*. 1998, Signal Processing Laboratory, Swiss Federal Institute of Technology at Lausanne.
9.El-Maliki, M. and A. Drygajlo. *Missing features detection and handling for robust speaker verification*. in *Eurospeech*. 1999. Budapest, Hungary.
10.Shao, Y. and D. Wang, *Robust speaker recognition using binary time-frequency masks*, in *ICASSP*. 2006.
11.Davis, G.M., *Noise Reduction in Speech Applications*. 2002, New York: CRC PRESS LLC.
12.Rose, P., *Forensic Speaker Identification*. Taylor & Francis Forensic Science Series, ed. J. Robertson. 2002, London: Taylor & Francis.
13.Dempster, A., N. Laird, and D. Rubin, *Maximum likelihood from incomplete data via the EM algorithm.* J. Roy. Stat. Soc., 1977. **39**: p. 1-38.
14.Duda, R.O. and P.E. Hart, *Pattern Classification and Scene Analysis*. 1973, New York: Wiley.
15.Bozkurt, B., L. Couvreur, and T. Dutoit, *Chirp group delay analysis of speech signals.* Speech Communication, 2007. **49**: p. 159-176.
16.Ortega, J., J. Gonzalez, and V. Marrero, *AHUMADA: A large speech corpus in Spanish for speaker characterization and identification.* Speech Communication, 2000. **31**: p. 255-264.
17.Drygajlo, A. and M. El-Maliki, *Speaker Verification in Missing Features Detection and Handling for Robust Speaker Verification*, in *EUROSPEECH'99*. 1999: Budapest, Hungary.
18.Reynolds, D.A., T.F. Quatieri, and R.B. Dunn, *Speaker Verification Using Adapted Gaussian Mixture Models.* Digital Signal Processing, 2000. **10**: p. 19 41.

AUTHORS

DayanaRibas González,Engineer in Telecommunications and Electronics, PhD Student, CENATAV,Havana City, Cuba, dribas@cenatav.co.cu.

José Ramón Calvo de Lara,Engineer in Telecommunications and Electronics,Doctor of Philosophy, CENATAV, Havana City, Cuba, jcalvo@cenatav.co.cu.