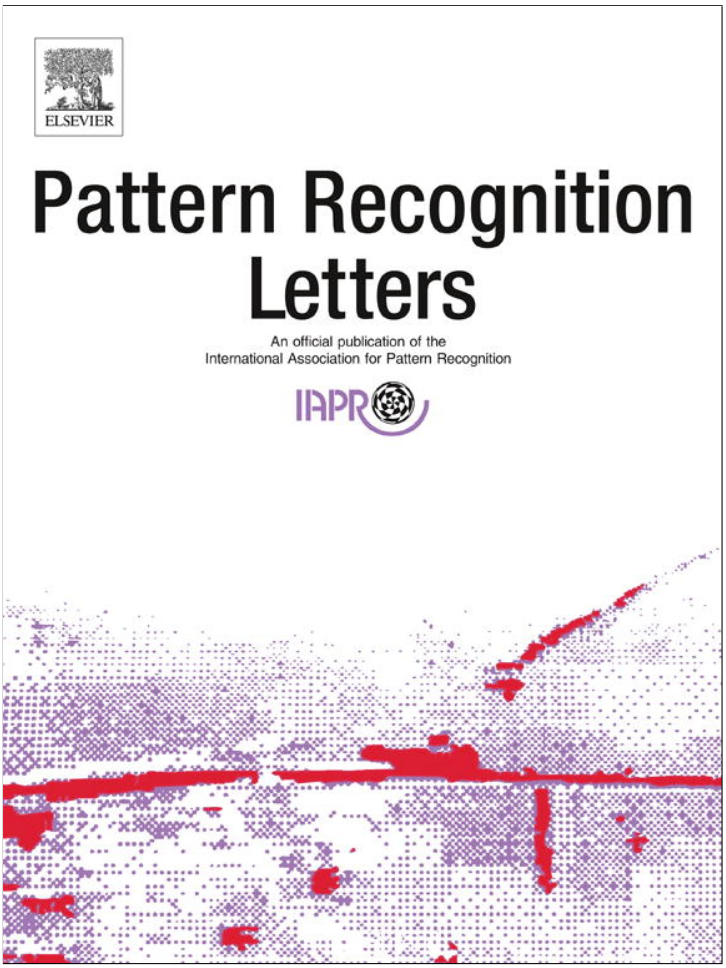

(This is a sample cover image for this issue. The actual cover is not yet available at this time.)

Pattern Recognition Letters 33 (2012) 2156–2165

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Is masking a relevant aspect lacking in MFCC? A speaker verification perspective

Jugurta Montalvão [a,*], Marcos Renato Rodrigues Araujo [b]

[a] Universidade Federal de Sergipe, 49100-000 São Cristóvão, Brazil
[b] Griaule Biometrics, Campinas, São Paulo, Brazil

ARTICLE INFO

Article history:
Received 24 November 2010
Available online 22 August 2012
Communicated by S. Sarkar

Keywords:
Mel-frequency cepstral coefficients
Ensemble interval histogram
Zero crossing with peak amplitude
Robustness
Spectral masking

ABSTRACT

We hypothesize that spectral masking may account for most of the gains in robustness against noise using ensemble interval histogram (EIH) and zero crossing with peak amplitude (ZCPA) compared to Mel-frequency cepstral coefficients (MFCCs). To test this hypothesis, we focus on this issue by comparing two MFCC implementations for which the only difference is spectral masking. The comparison involved biometric speaker verification tasks using two publicly available databases. The results confirm the superiority of MFCC with masking, thus corroborating our hypotheses that masking is a key aspect for improved robustness in feature extraction.

1. Introduction

Speaker identification or verification is a challenging biometric modality in which robustness remains an open question. Biometrics is the science of establishing the identity of an individual based on physical, chemical or behavioral attributes (Jain et al., 2007). The human voice carries several levels of information to the listener: a message through words, the language spoken, the emotional state and the identity of the speaker (Heck, 2001). It is produced by an effort made by the speaker, initiated in his brain and resulting in several neuromuscular commands and the voice itself (Rabiner and Juang, 1993), characterizing the human voice as a behavioral attribute. In this context, a biometric speaker recognition system is defined as a computer system capable of identifying a person based only on information delivered through his voice.

Mel-frequency cepstral coefficients (MFCCs) are frequently used as a low-dimensional set of features to represent short segments of speech. Since it was first conceived in 1974, MFCC has remained a powerful sound representation tool as it partly mimics human perception of sound color (Terasawa, 2009), and thus is popular in the signal processing community in almost its original form. MFCC is also applied to speaker verification. For instance, Ramos-Castro et al. (2007) extracted 19 MFCCs from overlapping short speech-signal frames and Hautamaki et al. (2008) used the 12 lowest MFCCs as acoustic features.

Different operating conditions during signal acquisition severely affect MFCC (e.g. headset type, channel response, background noise), which can lead to feature mismatch across training and recognition. To overcome this problem, most approaches retain MFCC as features but introduce some type of compensation. For example, Hautamaki et al. (2008) used cepstral mean normalization to remove linear channel distortion, along with RASTA filtering and feature warping to achieve robustness against channel and noise effects. According to the authors, "state-of-the-art text-independent speaker recognizers use mean subtraction at the utterance level, often referred to as cepstral mean subtraction (CMS)" (Hautamaki et al., 2008), even though CMS may degrade accuracy recognition of clean data (no channel mismatch).

Alternatively, noise compensation can be directly applied during MFCC computation through spectral subtraction per band and/or by changing the band logarithmic energy compression with constant root functions (possibly with adaptive root parameters). Nasersharif and Akbari (2007) compared four such strategies, including a new one proposed by the authors and found that adaptive root energy compression and noise compensation in sub-bands together outperformed all the others strategies in an isolated word recognition task.

There are many strategies for tackling the so-called session variability problem, roughly classified as: (a) feature domain compensation, such as cepstral mean normalization and RASTA; (b) score domain compensation, such as HNORM (Reynolds et al., 2000); and (c) model domain compensations, which includes subspace and factor analysis methods (Dehak et al., 2011; Kenny et al., 2007).

In the present study, a simple change in MFCC computation involving spectral masking is presented as a strategy for improving robustness. We compare MFCC to two main alternative features, the ensemble interval histogram (EIH) (Ghitza, 1994) and zero

* Corresponding author. Tel.: +55 79 88 76 17 36; fax: +55 79 21 05 66 84.
E-mail addresses: jugurta.montalvao@infonet.com.br, jmontalvao@ufs.br (J. Montalvão), marcos.araujo@griaulebiometrics.com (M.R. Rodrigues Araujo).

0167-8655/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2012.07.023

crossings with peak amplitude (ZCPA) (Kim et al., 1999). We show that these alternative features use spectral masking, which may explain their reduced sensitivity to external noise. The EIH technique is based on the assumption that there are dominant frequencies in limited bands of the signal (voice) spectrum, so the inverse level-crossing length accepts values around the dominant frequency, providing great robustness against noise, but making it strongly sensitive to the choice of levels in the level-crossing detector. ZCPA is an alternative approach in which the peak amplitude between adjacent zero-crossings is used as a nonlinear weighting factor for the corresponding frequency bin, providing a lower computational burden compared to EIH.

We hypothesize that the spectral masking effect, explained in Section 2, is the main reason behind the superiority of EIH and ZCPA. In Sections 3 and 4 we show how EIH and ZCPA use lateral masking, whereas MFCC does not. In Section 5, we propose a simple modification of the MFCC algorithm to allow for inclusion of the missing masking effect. The resulting improved robustness against additive white Gaussian noise (AWGN) is demonstrated in Section 6. Finally, biometric experiments on speaker verification in Section 7 clearly show an impressive performance gain, thus corroborating our initial hypothesis.

2. Spectral masking and auditory filters

Psychoacoustics is the study of subjective human perception of sounds. In this field of study, a well-known empirical observation is that in some situations, an otherwise clearly audible sound can be masked by another louder sound. This is called masking, a phenomenon that occurs because any loud sound distorts the absolute threshold of hearing.

Masking was defined by the American Standards Association in 1960 as "the process by which the threshold of audibility for one sound is raised by the presence of another (masking) sound" (Moore, 1989). Masking may be due to simultaneous sounds or separated sounds over time. In this paper, only simultaneous masking (i.e. a signal presented at the same time) is taken into account.

Simultaneous masking was used by Fletcher (1940) in experiments that laid the foundation for the concept of a critical band. In these experiments, he measured the threshold of a sinusoidal signal as a function of the bandwidth of a bandpass noise masker centered at the sinusoidal frequency. Fletcher observed that when the noise bandwidth was increased, the masking threshold also increased, as expected. However, he also observed that the noise bandwidth no longer influenced the masking threshold after a certain critical value, the critical band. Fletcher then suggested that the peripheral auditory system behaves as if it contains a bank of bandpass filters, with continuously overlapping center frequencies, so-called auditory filters. Clearly, a simplifying assumption was that the shape of the auditory filter could be approximated as a simple rectangle, with a flat top and vertical edges. Even though the original models proposed by Fletcher were inaccurate, and have been revised and adjusted in a myriad of more recent studies, the critical band and auditory filter concepts still pervade current research.

Experimental results obtained since 1940 corroborate the idea that a signal with bandwidth narrower than an equivalent rectangular bandwidth (ERB) excites a limited length of the basilar membrane (BM), and the absolute threshold of hearing inside this BM segment consequently increases. ERB is a function of a given central frequency, which can be approximated as (Moore and Glasberg, 1983):

$$\mathrm{ERB}(f) = 25 + 75(1 + 1.4f^2)^{0.69},$$

where ERB(f) is in Hz and f is the central frequency in kHz.

For instance, at 1 kHz the ERB approaches 162 Hz, whereas at 100 Hz it approaches 101 Hz. If we follow the path paved by Fletcher, the cochlea can be viewed as a filter bank consisting of bandpass filters whose bandwidths are roughly given by the corresponding ERB. In other words, an observer experiences an inability to resolve harmonic sounds whose frequency difference is smaller than a critical band. According to most broadly accepted physiological models of human ear processing, sounds are perceived through the mapping of air vibration to BM activity. Moreover, a critical band may form around any central frequency (i.e. the harmonic fundamental frequency of the excitation). Therefore, by arbitrarily choosing one frequency as a starting point, between 20 Hz and 16 kHz, 24 non-overlapping critical bands may be arranged side by side (Zwicker and Feldtkeller, 1981). Their central frequencies roughly follow a uniform distribution below 1 kHz, and a logarithmically increasing spacing above 1 kHz (Becchetti and Ricotti, 2004).

Given the dependence of ERB on the central frequency, the usual linear scale for frequency in cycles per second or Hz is frequently replaced with experimentally determined frequency scales, so that ERB is approximately constant on these scales. Here we use one of these scales, the Mel frequency, whose functional relationship with the scale in Hz is shown in Eq. 1. In other words, whenever we map frequency in cycles per second (Hz) to the new frequency axis (the Mel scale), all critical bands approximately exhibit the same bandwidth (in Mel).

$$f_{\mathrm{Mel}} = 2595 \log\left(1 + \frac{f_{\mathrm{Hz}}}{700\ \mathrm{Hz}}\right). \quad (1)$$

In spite of the rectangular shape first proposed by Fletcher for auditory filters, a different shape became more popular. This shape was obtained by simply inverting the psychophysical tuning curve, which is similar to the neural tuning curve (obtained as the level of a tone required to produce a fixed output from a single neuron) (Moore, 1989). This inverted shape can roughly be approximated by a triangle, as in typical MFCC implementation, but it can be more precisely obtained through the notch noise method proposed by Patterson, 1976, thus yielding the rounded-exponential filter shape (roex) function model $\mathrm{roex}(p,r) = (1-r)(1+pg)e^{-pg} + r$, where p describes the slope of the filter, r controls the filter dynamic range, and g is the frequency deviation from the filter center.

3. MFCC and masking

3.1. MFCC history and typical implementation

The signal analysis currently known as MFCC was first proposed by Bridle and Brown (1974) as the log spectrum transformed through a 19-channel filter bank, so that corresponding energies were in turn cosine-transformed into 19 "spectrum shape" coefficients. Mermelstein (1976) named this algorithm Mel-based cepstral parameters and use the MFCC acronym for the first time. In his work, he applied the algorithm to measure inter-word distances for a time-warping task in speech recognition.

Filter banks used in MFCC mimic the auditory critical-band filter bank, with fixed bandpass centers. The output energies from these filters can be regarded as a subsampling of the spectrum. The logarithms of these energies are taken to mimic loudness compression in mammalian ears. The output in MFCC computation is a low-dimensional representation of compressed bandpass filter energies, obtained through a discrete cosine transform (DCT).

The lower DCT projection coefficients used in MFCC roughly correspond to the principal component projections of speech signal spectra and good word recognition can be obtained using only a

few of these principal components (Pols, 1971). More precisely, Pols (1966) showed that the first six eigenvectors of the covariance matrix for Dutch vowels uttered by three speakers (from 17 bandpass filter energies) accounted for 91.8% of the total variance. Fortunately, the principal eigenvectors he found are very similar to the vectors used in DCT projection, which may explain why DCT roughly provides almost decorrelated cepstral projection, capturing the phonetically important characteristics of speech.

Even though MFCC has been regarded as one of the simplest auditory models, it remains a powerful sound representation tool since it provides a linear and orthogonal coordinate space for human perception of "sound color" (Terasawa, 2009). MFCC became widely popular in the signal processing community in its almost original form, with triangular bandpass filters simulated through the weighting of Fourier spectra, yielding the so-called Mel spectrum and then the cosine-transformed MFCC. Over time, MFCC extraction is performed once for every short-time signal frame. Within these frames of typically 20–40 ms (although (Davis and Mermelstein, 1980) used 12.8 ms instead), speech signals can be conveniently approximated by stationary random processes. By denoting a whole signal (digitized utterance) as $s(n),\ n = 1, 2, \ldots, N_s$, and the fth frame of $s(n)$ as $x(m;f) = s(n_f + m - 1)$, where $m = 1, 2, \ldots, M,\ M \ll N_s$ and n_f denotes the sample where frame f begins, we can summarize a typical MFCC extraction in frame f as follows

1. Pre-emphasized signal $x(m;f)$ is weighted, usually with a Hamming window, according to $x_w(m;f) = w(m)x(m;f)$, where $w(m) = 0.54 - 0.46\cos(2\pi(m-1)/M)$.
2. $x_w(m;f)$ is padded with zeros and FFT-transformed to $X_w(k;f)$.
3. Magnitude values of $X_w(k;f)$ are partially weighted and summed, thus simulating bandpass filters (triangular filters).
4. Log values of the resulting partial sums, K_t overlapping sums (K_t triangular filters), are arranged as a vector of energy values.
5. DCT is applied to this vector and the first 10–25 resulting coefficients are taken, the MFCCs.

Therefore, MFCC mimics two perceptual features: the frequency response of the BM, and the compressive nonlinearity of auditory nerve excitations.

A current MFCC implementation by Malcolm Slaney (publicly available at https://engineering.purdue.edu/~malcolm/interval/1998-010/) in MATLAB code uses 40 triangular bandpass filters, 13 linearly spaced, from 133.3 to 866.7 Hz, and 27 non-linearly spaced from 866.7 to 6398.5 Hz. Another typical implementation, as proposed by Paulus and Hornegger (2004), uses only 25 triangular filters. Seven of them are linearly spaced, centered at frequencies of 150–450 Hz, and have a constant bandwidth of 100 Hz. For the remaining 18 triangular filters, three groups of six each cover an octave between 500 and 4000 Hz (500–1000 Hz, 1000–2000 Hz, and 2000–4000 Hz).

Since the work of Bridle and Brown (1974) and Mermelstein (1976), the peripheral auditory system has often been modeled as a bank of overlapping triangular bandpass filters (Moore and Glasberg (1983)). However, there have been some exceptions. For instance, Hermansky (1990) and Zheng et al. (2001) used a piecewise shape to simulate a critical-band-masking curve to approximate the asymmetric masking curve, which exploits the proposal of Zwicker and Feldtkeller (1981) that auditory filter shape is approximately constant on the Bark scale. Indeed, although triangular filters are usually preferred, Zheng et al. (2001) presented many tests for several MFCC implementations, with three filter shapes (triangular, rectangular and Schroeder) implemented on two scales (Bark and Mel), with and without filter overlapping and with the number of filters ranging from 35 to 40. All these MFCC variants were compared in an automatic speech recognition task with a standard Mandarin database of utterances (the 863 database). The experimental results indicated that differences between the Bark and Mel scales and among the filter shapes tested are not significant (Zheng et al., 2001). However, the absence or presence of filter overlapping led to a significant difference.

3.2. Spectral masking in MFCC

The main link between masking and MFCC is rather straightforward: the MFCC filter bank approximately resembles the critical-band bank. Therefore, the roex-like shape of auditory filters, obtained through experiments with the masking effect, can roughly be approximated by triangles, as in typical MFCC implementation, where linear filters with a triangular bandpass shape are used.

Nonetheless, in spite of the clear masking-related origins of the filters used in MFCC, the masking effect is lost when a bank of linear bandpass filters is used to model nonlinear phenomena that occur in the peripheral auditory system. Indeed, even if the spectral details disappears, since sound energy inside the spectral interval covered by one bandpass filter is mapped to a single averaged scalar value, it is clear that all the small spectral details still affect this averaged energy per critical band, whereas the masking effect would rather conceal some small spectral contributions. It is worth highlighting that roex-shaped curves are hearing threshold contours, not linear weights, as in MFCC.

Another important difference between typical MFCC and the auditory model suggested by Fletcher (Section 2) is that the bank of bandpass filters would have continuously overlapping center frequencies whereby the auditory filters are not centered at prefixed frequencies. Thus, triangular filters with fixed centers may reinforce noise when these centers fall in parts of the spectrum where the noise is greater than the signal. This may explain in part why standard MFCC is so sensitive to noise, in addition to reasons already discussed by Wu and Cao (2005) and Nasersharif and Akbari (2007), according to which the log function used in MFCC computation tends to reinforce background noise.

4. Spectral masking in EIH and ZCPA

The EIH method of Ghitza (1994) has better robustness to additive Gaussian noise than Fourier-based methods. In the EIH method, the frequency content of a signal is estimated from the spiking behavior of simulated auditory nerve fibers, producing a frequency domain representation similar to a Fourier magnitude spectrum.

In the original study, Ghitza (1994) used a filter bank of 190 overlapping cochlear channels (bandpass filters), logarithmically spaced between 200 and 7000 Hz, followed by an array of five level-crossing detectors per filter output that simulate the auditory nerve fibers which innervate one inner hair cell. Only intervals between successive upward-going level crossings are considered and a histogram of the inverse interval (i.e. instantaneous frequency estimates) is computed. Two typical bin allocations for the interval histogram were proposed, the finest one corresponding to 128 bins linearly distributed from 0 to 4000 Hz. In this case, each bin covers a frequency interval of 31.25 Hz. For instance, given a bin centered at 100 Hz, every interval between 8.65 ms (1/115.6 Hz) to 11.85 ms (1/84.4 Hz) is taken as an entry to this bin. Compared to MFCC, the EIH method has two remarkable sources of robustness against noise:

- Level-crossing detectors estimate the dominant frequency present at the output of each bandpass filter. If a dominant frequency lies within the band, noise outside the band is filtered, increasing the signal-to-noise ratio (SNR) before level-crossing detection.
- Five level-crossing detectors, placed at different levels, extract redundant information from the signal within each band.

In EIH, the masking effect comes from the combination of frequency estimates from crossing intervals and frequency quantization (in histogram bins). As an illustration, we consider a "masking" signal given by $s_0(t) = \sin(2\pi 100t)$, which clearly is expected to fall in the histogram bin centered at 100 Hz. In the presence of a second signal, $s_1(t) = A\sin(2\pi(100 + \Delta F)t + \phi)$, the resulting signal $s(t) = s_0(t) + s_1(t)$, for small values of A ($A < 1$), may produce the very same histogram as $s_0(t)$, if interval perturbations due to $s_1(t)$ are not greater than 15.625 Hz. In this case, the presence of $s_1(t)$ is not "noticed" in the spectral histogram representation; in other words, it is masked by $s_0(t)$.

A more precise analysis of this masking effect can be presented as follows, for an upward-going zero-level crossing detector. Given A, the strongest interval perturbation occurs when extreme values of $s_1(t)$ ($\pm A$) perfectly cancel $s_0(t)$, thus maximally separating/approximating zero-level crossing instants, as illustrated in Fig. 1.

Therefore, the maximum period of $s_1(t)$ must be $T_0 \pm 2\Delta t$, where T_0 is the period of $s_0(t)$, so that $\sin\left(2\pi\frac{\Delta t}{T_0}\right) = A$. Consequently,

$$\Delta t = \frac{T_0 \arcsin(A)}{2\pi} \tag{2}$$

and

$$T_1 = T_0 \pm 2\Delta t, \tag{3}$$

where T_1 is the deviated interval. From Eqs. (2) and (3), the deviated period can be expressed in terms of T_0 as

$$T_1 = T_0\left(1 \pm \frac{\arcsin(A)}{\pi}\right). \tag{4}$$

It is easy to see that the strongest period perturbation is induced by an interfering harmonic at $F_I = 1/2T_1'$, with $T_1' = T_0\left(1 - \frac{\arcsin(A)}{\pi}\right)$, or odd harmonics of F_I. Considering that band filtering is applied to the signal before the level crossing detectors, we assume that only F_I and $3F_I$ are relevant in this analysis.

Fig. 2 presents the maximum frequency deviation as a function of F_I for $F_0 = 100$ Hz . It is worth noting that maximum deviation is obtained with F_I outside the band covered by that histogram bin. Consequently, bandpass filtering further reinforces the masking effect by attenuating interfering components far from 100 Hz. By contrast, for $F_0 = 1000$ Hz, as long as the bin bandwidth remains constant, the same deviation effect of 15.625 Hz is attained at 507.81 and 1523.43 Hz (with a much lower interfering amplitude of only $A = 0.048$).

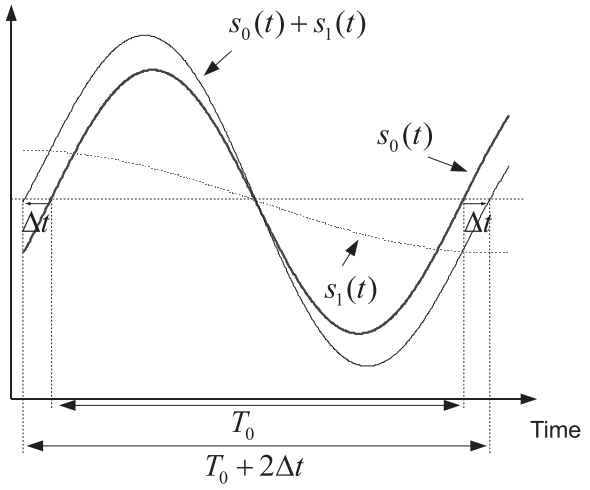

Fig. 1. Increase in the maximum zero-crossing interval due to a single harmonic interference.

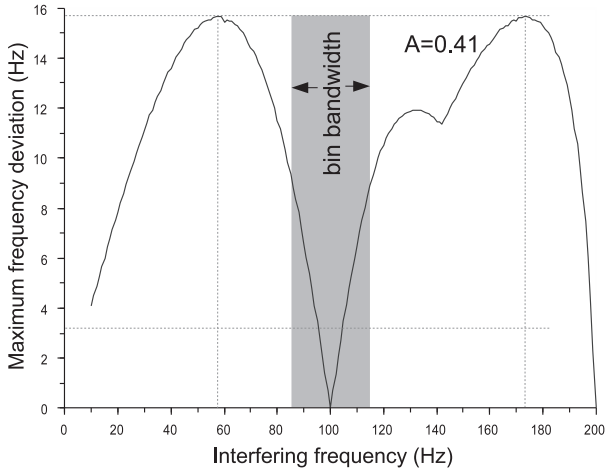

Fig. 2. Maximum ΔF for $A = 0.41$ and $F_0 = 100$ Hz (average values calculated numerically).

Indeed, the whole masking effect due to the joint effect of bandpass filtering and level crossing detection is frequency- and amplitude-dependent, and a deeper analysis is beyond the scope of this paper.

Unfortunately, EIH is strongly sensitive to the choice of levels in the level-crossing detector, since each level is differently affected by noise (Kim et al., 1999). To circumvent this difficult design problem, Kim et al. (1999) proposed an alternative method, ZCPA, in which the peak amplitude between adjacent zero-crossings is used as a nonlinear weighting factor for the corresponding histogram bin (frequency bin) for each bandpass filter. The use of a single zero-crossing detector per bandpass filter output provides a much lower computational burden, whereas its performance in recognition tasks is similar to that of EIH, and therefore is better than MFCC in the presence of noise (Kim et al., 1999).

If zero-crossings and histograms are used, the masking effect illustrated here for EIH can be extended to ZCPA. Kim et al. (1999) showed that a spectrum based on zero-crossing has a tendency to enhance a dominant signal component. Moreover, as illustrated here, a dominant signal component within a given frequency band may mask weaker components, even from outside this band, if zero-crossing-based frequency estimates lie within a limited histogram bin bandwidth.

We believe that this implicit masking effect for both strategies accounts for most of their performance gains in terms of robustness compared to MFCC. Nevertheless, we do not propose an alternative for either EIH or ZCPA. Instead, we focus on the missing aspect of traditional MFCC algorithms, the masking effect, and propose a straightforward way to include masking in an MFCC-like algorithm. Therefore, we do not compare the performance of our approach to that of EIH or ZCPA. Instead, we limit ourselves to experimental comparisons between MFCC with and without masking.

5. Proposed MFCC-like algorithm

As mentioned in Section 3, there is no consensus on practical MFCC implementation because there are many possible choices for filter shape and number. Even filter bandwidth and distribution along the frequency scale (with or without overlapping) are a matter of arbitrary choice.

Investigative studies such as that of Zheng et al. (2001) provide hints on possible choices. For instance, Zheng et al. (2001) showed that in an automatic speech recognition context, differences

between the Bark and Mel scales and different filter shapes are not relevant. By contrast, whether or not the filters overlap makes a big difference. The authors obtained the best classification results with 35–40 filters with approximately 158 Mel per filter.

In this section we use these results and further choose a slightly new approach to the usual MFCC. This new approach has the merit of allowing inclusion of a masking effect through a very simple algorithmic modification.

5.1. Alternative MFCC implementation

In our alternative MFCC implementation we replace the usual FFT-based analysis with a non-inversible DFT-like signal projection. This main change allows for masking implementation in a very simple way. Furthermore, other minor aspects of our MFCC implementation are as follows:

- short-time analysis: 25 ms per frame;
- overlapping between frames: 82% (advance of 4.5 ms per frame);
- blackman window instead of (typical) Hamming window;
- frequency scale: Mel;
- filter shape: triangular or rectangular;
- filter bandwidth: constant on the Mel scale; and
- frequency analysis: modified DFT, on the Mel scale, instead of FFT.

Given all overlapping short frames $x(m;f) = s(n_f + m - 1)$ (Section 3) from a signal $s(n),\ n = 1, 2, \ldots, N_s$, to discard silent frames (or frames with acoustic energy that is too low), we first estimate the variance of each Blackman window for $x(m;f)$ as:

$$v(f) = (1/(M-1))\sum_{m=1}^{M}(x(m;f)w_B(m) - \bar{s}(f))^2,$$

where $w_B(m) = 0.42 - 0.5\cos(2\pi(m-1)/M) + 0.08\cos(4\pi(m-1)/M)$ and

$$\bar{s}(f) = (1/M)\sum_{m=1}^{M}(x(m;f)w_B(m)).$$

Then we set an adaptive energy threshold:

$$L = \frac{\text{mean}(v(f)) + \min(v(f))}{2},$$

where $\text{mean}(v(f))$ and $\min(v(f))$ denote the mean and minimum values of $v(f)$, respectively, over all frames.

This threshold procedure is illustrated in Fig. 3 for a signal corresponding to the utterance (in Portuguese) "chocolate, zebra, banana, táxi" (approx. 3 s).

All remaining short-signal frames $x(m;f)$ are then subjected to frequency analysis. Typical MFCC implementations use FFT for computational efficiency reasons. Since we are including a masking effect on the nonlinear Mel scale, a more suitable approach is the use of a DFT-like transform matrix T whose rows are complex exponential vectors at non-linearly spaced frequencies. In other words, T is an $M \times K$ matrix, whose entries are

$$T(m,k) = \exp(-j(m-1)2\pi f_{\text{Hz}}(k)/F_s),$$

where F_s stands for the sampling frequency in samples per second, $m = 1, 2, \ldots, M,\ k = 1, 2, \ldots, K,\ j = \sqrt{-1}$ and

$$f_{\text{Hz}}(k) = 700(10^{f_{\text{Mel}}(k)/2595} - 1)\text{Hz}$$

and

$$f_{\text{Mel}}(k) = 150 + \frac{(2840-150)(k-1)}{(K-1)}\text{Mel}.$$

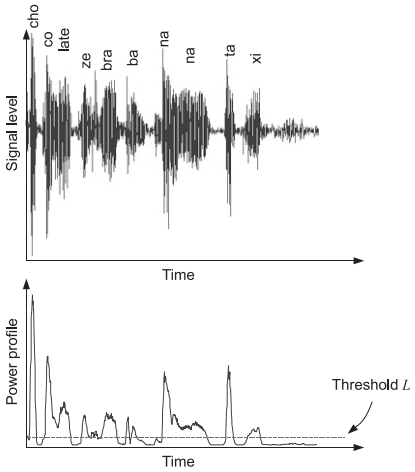

Fig. 3. First signal segmentation using the power profile: frames with a signal energy below the automatically set threshold L are discarded.

This DFT-like transform matrix is not square (and is therefore not inversible) if $K < M$. Moreover, rows are associated with frequencies (in Hz) that are mapped back from an equally spaced grid of frequencies on the Mel scale, and therefore yield a non-linear grid of frequencies in Hz. We arbitrarily choose $K = 145$, and the resulting correspondence between the perceptual Mel scale and the frequency in cycles per second (Hz) is illustrated in Fig. 4.

Unlike FFT, this choice is not based on computational efficiency; nevertheless, for short windows with 553 samples (25 ms at F_s = 22050 samples/s), the proposed analysis yields 80,185 complex multiplications. By contrast, to obtain the same minimum spectral interval between bins of 13.4 Hz with FFT, it would be necessary to increase the signal frames at low frequencies. Consequently, a computational burden of 17575 complex multiplications would be considered instead. In other words, our DFT-like analysis requires fewer than five times as many complex multiplications as the equivalent FFT, which is not prohibitive.

Matrix T is computed once and every overlapping frame for the samples is time-to-frequency mapped according to

$$X = T^t x_B,$$

where $x_B(m;f) = x(m;f)w_B(m)$ and t denotes matrix transposition.

Since $X(k;f)$ $(k = 1, 2, \ldots, K)$ denotes the spectrum of $x_B(m;f)$ $(m = 1, 2, \ldots, M)$ and k denotes linearly spaced frequencies on the Mel scale (thus non-linear on the Hz scale), linear filters can be implemented through constant-bandwidth windows, as illustrated in Fig. 5.

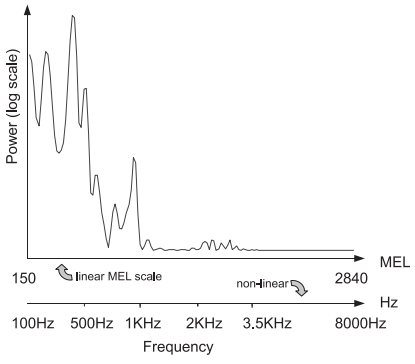

Fig. 4. Frequency analysis on the Mel scale and its corresponding scale in Hz. If we assume that the Mel scale is linear, from a perceptual perspective the equivalent scale in Hz is non-linear.

We experimented with two filter response shapes, triangular and rectangular, associated with the following spectral weighting functions:

$$h_t(k;k_c) = \begin{cases} 1 - \frac{2|k-k_c|}{BW}, & 2|k-k_c| < BW, \\ 0, & 2|k-k_c| \geqslant BW, \end{cases} \quad (5)$$

$$h_r(k;k_c) = \begin{cases} 1, & 2|k-k_c| < BW, \\ 0, & 2|k-k_c| \geqslant BW, \end{cases} \quad (6)$$

where k_c denotes a discrete central frequency and BW is the constant filter bandwidth (number of discrete values). We arbitrarily set $k_c = 1, 5, 9, \ldots, 145$, which leads to 37 filters. Note that if $BW/2 > 2$, then the filter bands overlap. We experimented with values of BW from 10 to 26, with strong overlapping between filters. It is worth noting that $BW = 10$ ($\sim$168 Mel) roughly corresponds to the filter bandwidth in the usual MFCC implementation. Moreover, according to Eqs. (5) and (6), the first filter was centered at 100 Hz, whereas the last filter was centered at 8000 Hz. In both cases, they were placed far from the spectral boundaries, 0 Hz and the Nyquist frequency.

To simulate the filter bank effect, magnitude values of $X(k;f)$ are weighted and summed to provide filter outputs. These outputs are log-transformed and arranged in a log-energy vector E. For triangular filters, this corresponds to the following operations:

$$E(k_c;f) = \log\left(\sum_{k=1}^{K} X(k;f) h_t(k,k_c)\right). \quad (7)$$

For rectangular filters, we just replace h_t with h_r in this equation.

Finally, MFCC coefficients are obtained through conventional DCT of the (column) vector corresponding to the values of E for a given frame f.

The whole algorithm for MFCC extraction of a short-time signal frame is illustrated in Fig. 6.

As stated before, this alternative algorithm allows for straightforward implementation of masking. Nevertheless, in spite of all the changes, the alternative algorithm is still equivalent to typical MFCC extractors in terms of the resulting coefficients. To illustrate this equivalence, Fig. 7 shows the same short signal frame processed with a publicly available algorithm and the proposed algorithm, both adjusted to produce 12 MFCCs. The small differences observed are mainly due to positioning of the triangular filter center. Filter shapes in the Slaney implementation are weighted by the inverse of the corresponding triangle area, but this is not necessary in our approach, because the non-uniform density of spectral bins (provided by the modified DFT) produces an equivalent effect with non-weighted triangular filters.

5.2. Modified MFCC: inclusion of the masking effect

To include the masking effect in our algorithm, we first replace $k_c = 1, 5, 9, \ldots, 145$ with $k_c = 1, 2, 3, \ldots, 145$. Therefore, we now have one vector $h_t(k;k_c)$ centered at every discrete frequency value.

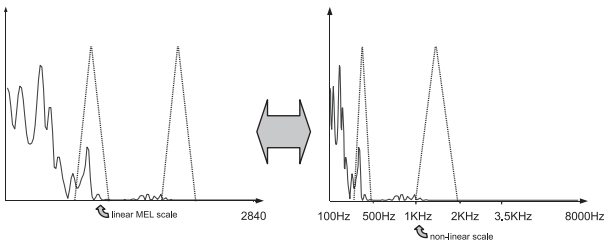

Fig. 5. Filter bank on the Mel and Hz scales.

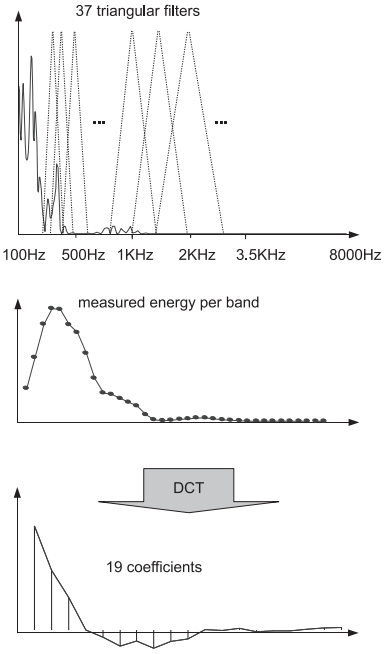

Fig. 6. MFCC extraction according to the usual approach through estimation of an energy vector in overlapping triangular windows (equally spaced on the Mel scale), followed by DCT of the log-transformed energy vector. Typically, the first coefficient (out of the 20 lowest DCT coefficients in this case) is systematically discarded.

This is motivated by the observation that a critical band may be formed around any central frequency (Fletcher, 1940; Zwicker and Feldtkeller, 1981). Moreover, we do not sum the weighted energies around k_c with $h_t(k;k_c)$, as in MFCC. Instead, we find the maximum weighted energy value for each weighting vector $h_t(k;k_c)$. Then we gather the indices for all the selected peaks (one per frequency center, k_c) in a vector $p(k_c)$ (note that the maximum value is not taken into account, since only its position or frequency is relevant):

$$p(k_c) = \arg\max_k (X(k;f) h_t(k,k_c)).$$

Vector p is a sequence of pointers to the discrete frequencies at which peaks occur, and some frequency values may be identified many times. In other words, the greater a spectral peak, the more likely it is that its position will be identified by p. The next step is to map p to a histogram h in which each bin $h(k)$ counts how many times frequency k was identified as a spectral peak in p.

Histogram h is closely related to the ensemble histogram in EIH (Ghitza, 1994) and plays the role of spectral representation. The simple modifications made to the MFCC extractor can be regarded as a sliding window (instead of fixed windows) from which energy peaks are taken, and all remaining spectral energy is discarded (masked) for each position of the sliding window, as illustrated in Fig. 8.

Finally, because our main goal is to test the effect of masking in an MFCC-like algorithm with as few changes as possible, we apply DCT to histogram h in the same way as log-filter bank energies are processed in MFCC.

6. Masking and robustness in AWGN

In this section, we empirically test how masking improves MFCC robustness against AWGN. In our tests, we arbitrarily selected a single zero-mean short signal frame (25 ms) corresponding to a male speaker uttering the vowel/a/ in a clean environment

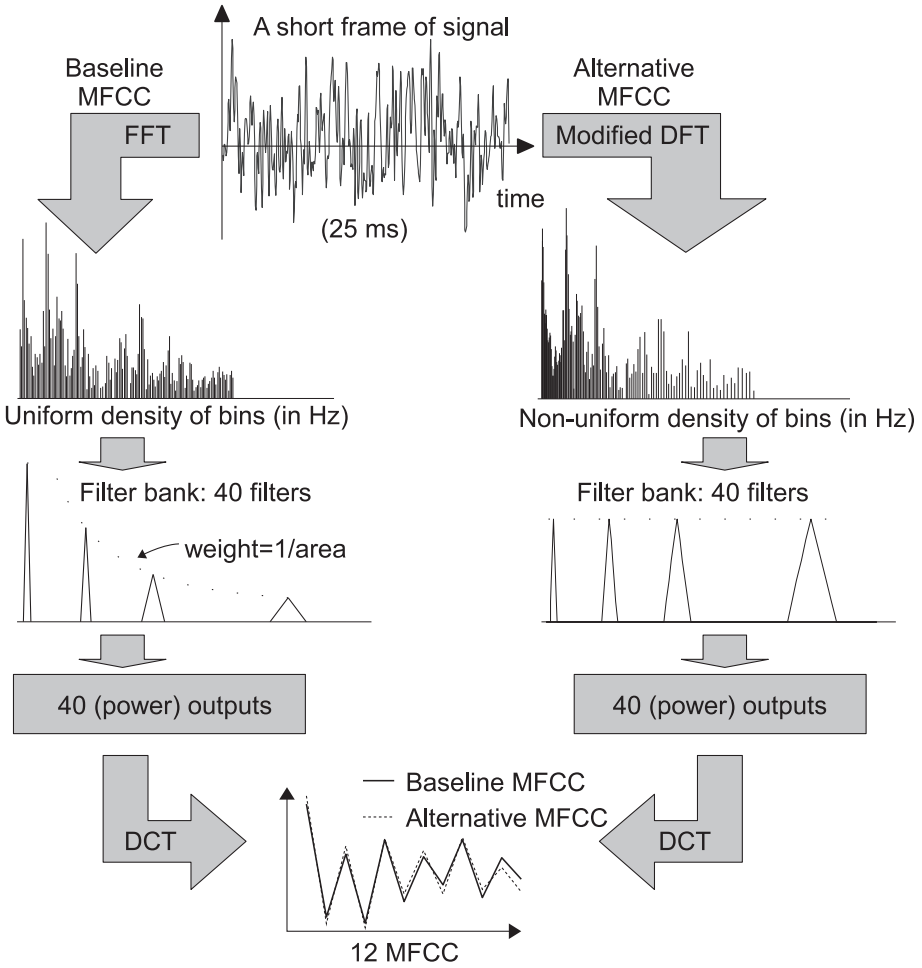

Fig. 7. Twelve MFCCs obtained using a typical implementation (Malcolm Slaney's Auditory Toolbox) compared to 12 MFCCs obtained with the alternative implementation.

(a noiseless laboratory). This short frame of clean signal $s_f(m)$ $(m = 1, 2, \ldots, M)$ was projected onto two feature spaces corresponding to: (a) 19 MFCCs, according to Section 5.1 and (b) 19 modified MFCCs including the masking effect, according to Section 5.2. Then a noisy version of the same short signal frame was obtained, $sn_f(m) = s_f(m) + n(m)$, where $n(m)$ denotes independently synthesized zero-mean white Gaussian noise. Moreover, the noise variance was adjusted to provide suitable SNR according to

$$\mathrm{SNR}(dB) = 10\log_{10}\frac{\sum_{m=1}^{M} s_f^2(m)}{\sum_{m=1}^{M} n^2(m)}.$$

Finally, the mean squared relative deviation (i.e. noisy coefficient minus clean coefficient divided by clean coefficient) between each pair of coefficient vectors was computed in 100 independent runs per method per noise level. The resulting average deviations are presented in Table 1 for triangular filters with a constant bandwidth of 168 Mel.

Since the signal samples, pre-processing, noise level, filter bandwidth and filter shapes are the same for both methods, differences in performance are expected to come from the masking effect implemented in the modified MFCC. At a low noise level (30 dB), MFCCs with and without masking are roughly equivalent in terms of sensitivity to white noise, whereas MFCC with masking is clearly less sensitive to noise for SNR $\leqslant$ 20 dB.

Experimental results presented in Section 7 suggest that the robustness of the modified MFCC method can be improved through changes in the filter shape and bandwidth. Indeed, because most results in the last part of Section 7 come from a specific choice of these parameters (flat filters instead of triangular with a wider bandwidth of 370 Mel), we also tested its relative deviation for different SNR values, which yielded the results presented in Table 2.

These better results are in agreement with the error rates presented in Section 7.

7. Experimental comparisons

7.1. Experiments using the BioChaves database

We performed biometric verification experiments using a publicly available database (The BioChaves database, available at www.biochaves.com/en/download.htm). Speech samples in this database correspond to signals recorded during the uttering of a single set of four words in Portuguese, chocolate, zebra, banana and táxi, that have identical English spellings, apart from táxi. The duration of each utterance was approximately 3 s and each subject uttered this set of words 10 times, five times during a first session and five times during a second session, approximately 1 month later, using a conventional headset (electret microphone and headset). Speech signals were digitized and recorded at 16 bits per sample, at 22050 samples per second. All recording was done under low background noise, although the noise environment was not controlled. Ten subjects took part in the experiment.

The database is small for biometric verification. However, our goal was to compare different versions of MFCC-based algorithms. We obtained a total of 1375 distinct pairwise comparisons

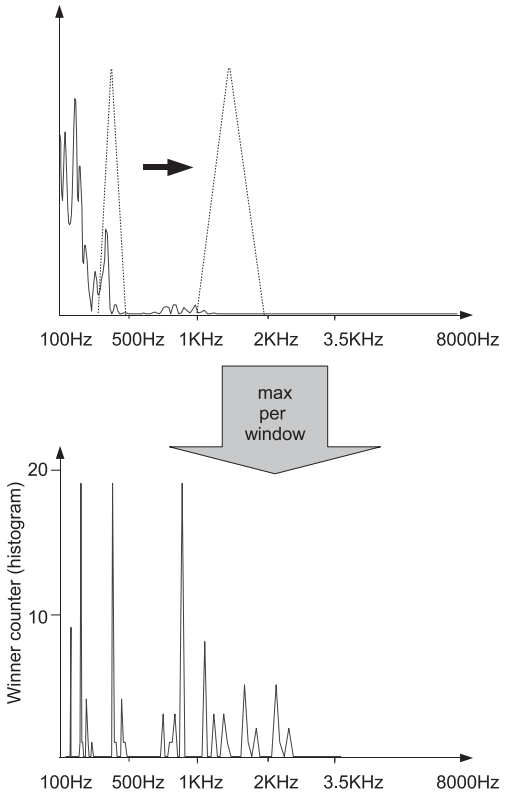

Fig. 8. Masking effect using a sliding filter on the Mel scale and a competition (max) operation.

Table 1
Relative deviation for different SNR values for triangular filters with a bandwidth of 168 Mel.

	30 dB	20 dB	10 dB	0 dB	−10 dB
MFCC	0.8	2.0	2.8	3.6	4.2
Modified MFCC	1.3	1.5	2.1	2.4	2.4

Table 2
Relative deviation for different SNR values for flat filters with a bandwidth of 370 Mel.

	30 dB	20 dB	10 dB	0 dB	−10 dB
Modified MFCC	0.1	0.4	1.2	1.5	1.7

between utterances from different sessions (1125 from different speakers and 250 from the same speaker), which was sufficient to show a clear and consistent difference between MFCC with and without masking.

It is worth noting that pairwise comparisons simulate a verification protocol with one utterance per enrollment and one per interrogation. Thus, given the short duration of each utterance for biometric purposes (only 3 s), we should expect high error rates compared to typical biometric experiments. However, this is not relevant to the comparison of algorithm performance.

Moreover, we did not take advantage of the fact that subjects utter the very same sentence to improve biometric results (e.g. through HMM or DTW). Instead, we just gathered extracted short-time features from each utterance as a set of randomly generated vectors (one set per utterance) and compared these sets using a K-nearest neighbors (K-NN) classifier. We arbitrarily chose $K = 5$ and slightly modified the classifier to obtain scores between 0 and 1 instead of average distances between feature vectors.

Therefore, in each experiment, every single utterance from one session (half of the database) was taken once as a prototype

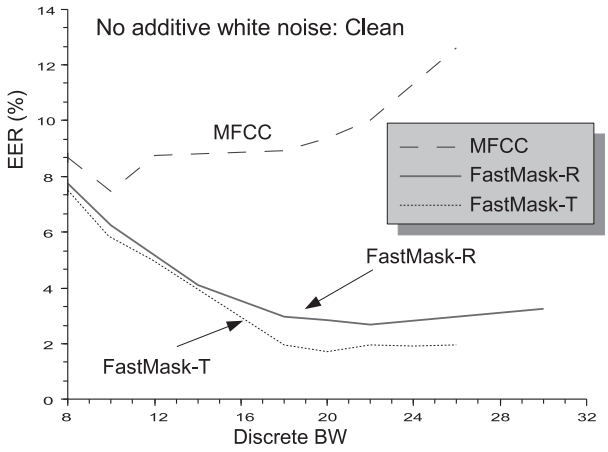

Fig. 9. Comparison of MFCC, FastMask-T and FastMask-R for a clean signal.

(enrollment) and then cross-compared to every other utterance from another session. When both compared utterances came from the same person, the resulting score was labeled as 'T' (true claimed identity), otherwise, it was labeled as 'F'. Thus, each MFCC algorithm yielded 250 scores labeled 'T' and 1125 labeled 'F'. Then the threshold was adjusted to minimize the equal error rate (EER)[1] for each set of scores and each MFCC algorithm.

The algorithms compared here are referred to as

(a) MFCC: 19 MFCCs per time frame;
(b) FastMasking-T: 19 MFCCs with masking using a triangular window; and
(c) FastMasking-R: 19 MFCCs with masking using a rectangular window, where the prefix 'Fast' conveys the idea that a sliding window for a short-frame signal spectrum has a lower computational burden than other masking implementations, typically based on actual filter banks in the time domain and with many parallel convolutions, such as in the EIH (Ghitza, 1994) and ZCPA (Kim et al., 1999) methods.

Fig. 9 presents EER values for clean sound (no additive white noise) and various discrete bandwidths. For BW = 10 (∼168 Mel), the conventional MFCC yields its best performance with EER ≈ 7.5%. It is interesting that 168 Mel is close to typical values found in practical MFCC applications. Nonetheless, the best performance was obtained with FastMask-T, with EER ≈1.7% for BW = 20 (∼337 Mel), closely followed by FastMask-R. It is evident that neither masking method is very sensitive to BW for values greater than 18 (304 Mel).

To test performance degradation under AWGN, independent noise was artificially added to clean utterances used for the test (the enrollment signal was kept clean). As shown in Fig. 10, for SNR = 10 dB, FastMask-T no longer outperformed FastMask-R, whereas MFCC remained the worst method, as expected.

Experimentation with other AWGN levels revealed that the FastMask-R method is more robust than FastMask-T, providing better results with a discrete BW ranging from 23 (387.82 Mel) to 25 (421.55 Mel). Fig. 11 gives a comparative overview of the performance for BW range three noise scenarios.

7.2. Experiments with the Ynoguti database

To provide statistically meaningful conclusions, a second set of experiments was performed using a larger corpus. This publicly

[1] The point at which the false alarm rate (FAR) equals the false rejection rate (FRR).

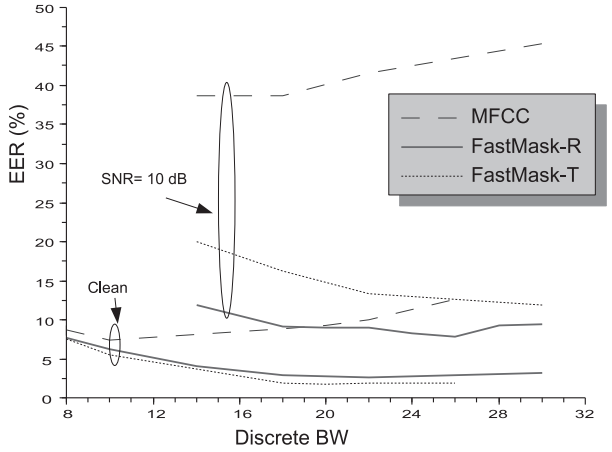

Fig. 10. Comparison of MFCC, FastMask-T and FastMask-R for SNR of 10 dB.

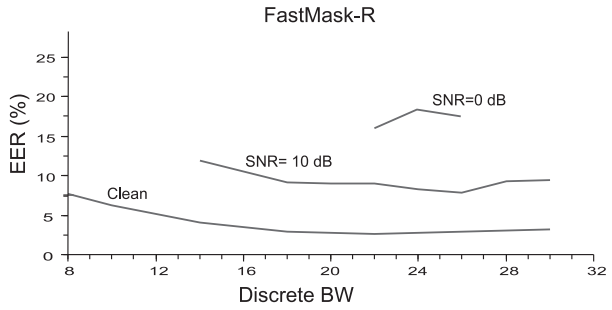

Fig. 11. Comparison of FastMask-R performance under three noise scenarios: clean, 10 dB and 0 dB.

available database, described by Ynoguti and Violaro (2008), includes voice samples from 40 speakers, each of whom utters between four and eight different sentences in Brazilian Portuguese.

Moreover, besides white synthetic noise, real-world noise samples were included in these experiments. We used three long sound files from NOISEX-92 (Varga and Steeneken, 1993): babble noise, car (Volvo) noise and factory noise, each of which was 235 s long.

Unlike the experiments in which we used a simple K-NN classifier to highlight the effects of masking, with very short training and testing signals (only ~3 s each), we experimented with longer training (~60 s) and test (~6 s) signals and a Gaussian mixture model (GMM)-based recognizer, adapted through likelihood maximization (via the Expectation–Maximization algorithm), closely following the approach of Reynolds and Rose (1995).

For each targeted user and for each SNR, we randomly selected 20 clean utterances (20 files), corresponding to approximately 60 s of signal to train a 50-component GMM with nodal variance (Reynolds and Rose, 1995). Then two other files (~ 6 s of signal) from a given user of the same gender were randomly selected and added to a weighted random 6-s segment of noise signal. More precisely, every noise signal was a randomly selected segment of the long noise files. The noise weight was adjusted to provide a specified SNR. Finally, the likelihood of the noisy test signal was computed using the trained GMM to produce a single score. This resulting single score was labeled as 'T' or 'F', according to whether the speaker in the test files was the same as in the training files or not.

This procedure was independently repeated 1175 times per target speaker for each SNR, which led to 400 'T' and 775 'F' scores. Moreover, in tests with a clean signal (no noise), in which the error rates were lower, we increased the number of tests to 3060 (1000 'T' scores and 2060 'F' scores). This protocol was used for MFCC without masking and the usual triangular filters and FastMask-R with a filter BW of 370 Mel (our choice to provide a suitable trade-off between performance with a clean signal and robustness against noise). As in the first set of experiments, we adjusted the threshold to find EER for each set of 1175 or 3060 scores. These EER are shown in Tables 3 and 4 for male and female speakers, respectively.

Table 3
EER (%) for male speakers.

	Clean	10 dB	0 dB	−10 dB	Noise source
MFCC		24.1	48.8	~50	White
	1.7	46.7	47.8	~50	Babble
		24.5	34.6	43.5	Volvo
		49.5	49.7	49.7	Factory
FastMask-R		22.5	35.5	43.5	White
	0.6	4.5	19.9	44.4	Babble
		1.0	2.2	17.8	Volvo
		9.2	20.3	41.4	Factory

Table 4
EER (%) for female speakers.

	Clean	10 dB	0 dB	−10 dB	Noise source
MFCC		23.6	49.6	~50	White
	3.1	42.7	46.8	~50	Babble
		25.2	37.1	43.7	Volvo
		49.2	~50	~50	Factory
FastMask-R		24.0	31.4	43.3	White
	1.9	8.7	19.0	43.1	Babble
		2.5	5.9	26.0	Volvo
		10.9	26.2	42.7	Factory

No noise compensation or reduction methods were used in any of the experiments, so any resistance to noise must come from masking. Strong sensitivity of the usual MFCC to additive noise is evident. For instance, apart from White and Volvo noise, even SNR of 10 dB is enough to increase EER above 40%. By contrast, for the very low-frequency Volvo noise, FastMask-R resisted a very strong noise level corresponding SNR of −10 dB.

8. Discussion and conclusions

We hypothesized that much of the effort in channel compensation and sophisticated pattern recognizer design can be saved by a very simple change in MFCC computation that involves inclusion of spectral masking in the algorithm. To test this hypothesis, we used a slight MFCC modification to allow for inclusion of masking in which averaging is replaced by maximization in the resulting algorithm.

In spite of its well-known drawbacks, MFCC is a popular and widely used feature for sound representation. This is the main reason why we did not propose a new feature. Instead, we demonstrated that the main MFCC drawback can be easily overcome with a slight algorithm modification involving inclusion of a masking effect. MFCC with and without masking were used for speaker verification. A performance gain was obtained when masking was used, mainly under strong noise conditions.

Since the only difference between the algorithms we compared is the masking effect (all algorithms were implemented with exactly the same parameters and the same DFT-like analysis), we conclude that this performance gain is due to the masking effect.

We also observed that rectangular (piece-wise) masking windows give better results under strong AWGN. Moreover, the best bandwidth was approximately twice the critical band.

All experiments were carried out using a public database. To allow further comparisons between the results reported here and the performance of other approaches for the same database, samples used in this work are freely available to download at http://www.biochaves.com/en/download.htm.

In conclusion, we confirmed that masking is a pivotal issue in robustness. A natural follow-up to this work will be the realization of more tests with different types of non-white noise. We only studied the effect of spectral masking in short signal frames, so we would expect even more interesting results when considering masking over time.

Acknowledgments

This work was supported by grants from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) to J.M. The work represents a collaboration between two Brazilian organizations, the Federal University of Sergipe and Griaule Biometrics, in an effort to produce a robust speaker verification system.

References

Becchetti, C., Ricotti, L.P., 2004. Speech Recognition. John Wiley & Sons, London.
Bridle, J.S., Brown, M.D., 1974. An Experimental Automatic Word-Recognition System. JSRU Report No. 1003, Joint Speech Research Unit, Ruislip.
Davis, S.B., Mermelstein, P., 1980. Comparison of Parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. 28, 357–366.
Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P., 2011. Front-end factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process. 19, 788–798.
Fletcher, H., 1940. Auditory patterns. Rev. Mod. Phys. 12, 47–65.
Ghitza, O., 1994. Auditory models and human performance in tasks related to speech coding and speech recognition. IEEE Trans. Speech Audio Process. 2, 115–131.
Hautamaki, V., Kinnunen, T., Franti, P., 2008. Text-independent speaker recognition using graph matching. Pattern Recognition Lett. 29, 1427–1432.
Heck, L.P., Reynolds, D.A., 2001. Speaker verification: From research to reality. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, Utah, 2001.
Hermansky, H., 1990. Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Am. 87, 1738–1752.
Jain, A.K., Flynn, P., Ross, A.A., 2007. Handbook of Biometrics. Springer-Verlag, New York.
Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P., 2007. Speaker and session variability in GMM-based speaker verification. IEEE Trans. Audio Speech Lang. Process. 15, 1448–1460.
Kim, D.-S., Lee, S.-Y., Kil, R.M., 1999. Auditory processing of speech signals for robust speech recognition for real-world noisy environments. IEEE Trans. Speech Audio Process. 7, 55–69.
Mermelstein, P., 1976. Distance measures for speech recognition, psychological and instrumental. In: Chen, C.H. (Ed.), Pattern Recognition and Artificial Intelligence. Academic Press, New York, pp. 374–388.
Moore, B.C.J., 1989. An Introduction to the Psychology of Hearing, third ed. Academic Press, New York.
Moore, B.C.J., Glasberg, B.R., 1983. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. J. Acoust. Soc. Am. 74, 750–753.
Nasersharif, B., Akbari, A., 2007. SNR-dependent compression of enhanced Mel sub-band energies for compensation of noise effects on MFCC features. Pattern Recognition Lett. 28, 1320–1326.
Patterson, R.D., 1976. Auditory filter shapes derived with noise stimuli. J. Acoust. Soc. Am. 59, 640–654.
Paulus, D.W.R., Hornegger, J., 2004. Applied pattern recognition. Algorithms and Implementation in C++,, Vieweg, Berlin.
Pols, L.C.W., 1966. Spectral Analysis and Identification of Dutch Vowels in Monosyllabic Words. Doctoral dissertation, Free University, Amsterdam.
Pols, L.C.W., 1971. Real-time recognition of spoken words. IEEE Trans. Comput. 20, 972–978.
Rabiner, L., Juang, B.H., 1993. Fundamentals of Speech Recognition. Prentice-Hall, Englewood Cliffs, NJ.
Ramos-Castro, D., Fierrez-Aguilar, J., Gonzalez-Rodriguez, J., Ortega-Garcia, J., 2007. Speaker verification using speaker- and test-dependent fast score normalization. Pattern Recognition Lett. 28, 90–98.
Reynolds, D.A., Rose, R.C., 1995. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. Speech Audio Process. 3, 72–83.
Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker verification using adapted Gaussian mixture models. Dig. Signal Process. 10, 19–41.
Terasawa, H., 2009. A Hybrid Model for Timbre Perception: Quantitative Representations of Sound Color and Density. Ph.D. thesis, Stanford University, Stanford, CA.
Varga, A., Steeneken, H.J.M., 1993. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun. 12, 247–251.
Wu, Z., Cao, Z., 2005. Improved MFCC-based feature for robust speaker identification. Tsinghua Sci. Technol. 10, 158–161.
Ynoguti, C.A., Violaro, F., 2008. A Brazilian Portuguese speech database. In: Proceedings of the XXVIth Brazilian Symposium on Telecommunication (SBrT'08), Rio de Janeiro, Brazil, 2008.
Zheng, F., Zhang, G.L., Song, Z.J., 2001. Comparison of different implementations of MFCC. J. Comput. Sci. Technol. 16, 582–589.
Zwicker, E., Feldtkeller, R., 1981. Psychoacoustique –L'Oreille, Récepteur d'Information. Masson, Paris.