Noname manuscript No.
(will be inserted by the editor)

Auditory-inspired Morphological Processing of Speech Spectrograms

Applications in Automatic Speech Recognition and Speech Enhancement

Joyner Cadore · Francisco J. Valverde-Albacete · Ascensión Gallardo-Antolín · Carmen Peláez-Moreno

Received: date / Accepted: date

Abstract A set of auditory-inspired features for speech processing is presented in this paper. The proposed acoustic features combine spectral subtraction and two-dimensional non-linear filtering techniques most usually employed for image processing. In particular, morphological operations like erosion and dilation are applied to a noisy speech spectrogram that has been previously enhanced by a conventional spectral subtraction procedure and filtered by an auditory filterbank. These methods had been applied both to speech enhancement and recognition. In the first application, anisotropic structural elements on gray-scale spectrograms have been found to provide a better perceptual quality than isotropic and reveal themselves as more appropriate for retaining the speech structure while removing background noise. A number of perceptual quality estimation measures have been employed for several Signal-to-Noise Ratios on the Aurora database. For speech recognition, the combination of spectral substraction and auditory-inspired morphological filtering has been found to improve the recognition rate in a noisy Isolet database.

Keywords Spectral subtraction · Spectrogram · Morphological processing · Image filtering · Automatic Speech Recognition · Speech Enhancement · Auditory-based features

J. Cadore
Universidad Carlos III de Madrid
Avda. de la Universidad 30, 28911 Madrid, Spain.
E-mail: jcadore@tsc.uc3m.es

F.J. Valverde-Albacete
E-mail: fva@tsc.uc3m.es

A. Gallardo-Antolín
E-mail: gallardo@tsc.uc3m.es

C. Peláez-Moreno
E-mail: carmen@tsc.uc3m.es

1 Introduction

Machine performance in Speech Recognition tasks is still far from that of humans [4].

The standard model conceptualizes the recognition process as a virtual channel in which word articulations (issued from the speaker's mind) are introduced and word percepts (selected among the receiver's hypotheses) are recognized. Data-driven statistical modeling techniques have made speech technologies lift off, enabling the introduction of these technologies into a variety of every day's applications.

With increased exposure to the public and the rise of commercial expectations, the sophistication and improvement of Speech Technology relies more and more on the Machine-Learning capabilities of computers. But the emphasis on Machine-Learning algorithms often comes at the expense of genuine insight into the nature of Spoken Language [16], which has brought about an ever-widening divide between the interests of phoneticians and linguists, on the one hand, and speech technologists, on the other. Yet, the suspicion looms above the latter crowd that the performance of some technologies stemming from this model is plateauing (Speech Recognition performance, notably) and the field is in need of a new technological paradigm.

In ASR (Automatic Speech Recognition), the standard procedure begins with the encoding of the speech signal as a string of feature vectors (also called *acoustic frames*) [36]. This is followed by an acoustic decoding stage where a maximum-likelihood (ML) strategy is applied to a set of generative models (e.g. the Hid-

den Markov Models, HMM) and the feature vectors to obtain a basic speech-unit percept correlate. Finally, the linguistic decoding process brings to bear the lexical and linguistic resources needed to obtain the word sequences [25, 1].

Among non-satisfactorily tackled challenges in ASR are various types of speech variability (rate, accent, dialect, gender, etc.) and background noise. One way to address these limitations is trying to imitate human acoustic capabilities, e.g. finding a more suitable auditory model.

In this paper we focus on the noise problem where humans are known to perform remarkably well whilst machines still lag behind [36]. There are many solutions inspired by the human auditory system aimed at solving this issue, e.g. feature extraction based on the well-known mel-frequency cepstral coefficients (MFCC, [7]), and on the Gammatone-based coefficients (GTC, [41]). Other solutions use the so-called *spectro-temporal features*, that consider both the time and frequency domains in the feature extraction stage ([32, 19]).

Following that line of work, in this paper we use morphological filtering over the spectrogram of a noisy signal to mimic some properties of the Human Auditory System (HAS), such as frequency and temporal masking, thereby filtering out as much noise as possible [9].

Morphological filtering is an image-processing tool for extracting image components that are useful for purposes like thinning, pruning, structure enhancement and noise filtering [15]. Since our goal is to emphasize the areas of interest of the spectrogram, the structural element shapes should be based on the spectro-temporal masking properties of the HAS, as in [9]. Our approach consists on substituting binary structural elements—thus avoiding the need for thresholding—by full gray-scale ones and we propose anisotropic structural elements as a means to better characterize the response of the ear.

Previous to the application of these ideas in ASR taks, we present some experiments for speech enhancement in order to assess the benefits of morphological filtering. Spectral Subtraction (SS, [3]) is used as a preprocessing stage to subsequently apply morphological filtering on the spectrogram thus producing a perceptually enhanced signal where *musical* noise (mainly caused by SS) has been reduced. In order to thoroughly evaluate the filtered signals we use a large quantity of speech utterances which, on the other hand, precludes the use of subjective quality measures. For this reason, estimations of these subjective opinions are computed from a set of objective quality measures [21, 22, 38, 2].

Finally, we apply this method on ASR tasks. In particular, we have have applied morphological filtering on a cochleogram with structural elements designed to model the HAS masking effects. This new features had been tested on a hybrid MLP/HMM recognizer.

This paper is organized as follows: section 2 introduces the main notions of the HAS employed in this paper in section 3, we present the preprocessing stage, spectrogram calculation and spectral subtraction. Section 4 is devoted to the explanation of our vision of morphological processing to imitate HAS capabilities. Finally, in sections 5 and 6 we describe two applications of the proposed method (Speech Enhancement and ASR, respectively) and end with some conclusions and ideas for future work in section 7.

2 Speech Processing at the Cochlear Level

2.1 Sound level

The basic quantity over which perception of sound is measured is *sound pressure level* which is a normalized, logarithmic[1] sound pressure:

$$L_p = 20 \log \frac{p}{p_0} (\mathrm{dB\,SPL}) \tag{1}$$

where p is sound pressure and $p_0 = 20\ \mu Pa$ is the reference sound pressure, the lowest audible pressure for human ears at mid-frequencies.

Another quantity is *sound (intensity) level*, a normalized, logarithmic intensity level:

$$L_I = 10 \log \frac{I}{I_0} (\mathrm{dB\,SL}) \tag{2}$$

where $I \propto p^2$ is *acoustic intensity*, an energy related quantity. When using $I_0 = 10^{-12} N/m^2$ for reference, both levels can be equated and we drop the subindex:

$$L = 20 \log \frac{p}{p_0} (\mathrm{dB\,SPL}) = 10 \log \frac{I}{I_0} (\mathrm{dB\,SL}) \tag{3}$$

Both dB SPL and dB SL could, in this case, be simple notated as dB.

2.2 Auditory filterbank description

It is widely accepted that the cochlea carries out a logarithmic compression of the auditory range whereby higher frequency intervals are represented with less detail than lower frequency ranges. This realisation stems from experiments to detect critical bands, which is the frequency bandwidth around a center frequency whose

[1] Unless otherwise noted log refers in this paper to base-10 logarithms.

components affect the sound level and pitch perception of the center frequency.

In this light, the notion of an auditory filterbank relates three concepts:

- A discretization of a frequency range into N bands.
- A choice of the center of the bands to be related to special frequencies or frequency ranges in the inner ear, which entails the definition of a *frequency scale*.
- A choice of the bandwiths and shapes of the different filters that takes into consideration the notion of *critical bands*.

The use of logarithmic frequency scales eases the conceptualization of phenomena like masking (see Figure 1), and we will consider several scales of logarithmic frequency: *Mel*, *Bark* and the *Equivalent Rectangular Bandwidth -induced (ERB) scale*. All of them use methods to calculate the critical bandwidths at different center frequencies and at the same time define scales of equal difference in perception of pitches/levels related to those center frequencies.

2.2.1 The MEL scale

The Mel scale is a very well-known logarithmic transformation of the frequency scale due to Stevens, Volkman and Newmann in 1937 [40]:

$$F_m(f) = 2595 \log(1 + \frac{f}{0.7}) \tag{4}$$

m in mel and f in kHz. This frequency transformation is in the core of the most popular feature extraction procedure in ASR: the Mel Frequency Cepstral Coefficients (MFCC) where a filterbank of triangular overlapping filters uniformly distributed in the mel scale is usually employed. This is one of our choices for testing our thesis as will be explained in section 6.

2.2.2 Critical band and critical-band rate scale

The Bark scale was first defined by Zwicker and colleagues but the actual formulas for transforming from linear frequency to bark scale have been taken from [44, 17][2]:

$$F_z(f) = 13 \cdot \arctan(0.76 \cdot f) + 3.5 \cdot \arctan(f/7.5) \tag{5}$$

with critical band rate, z in bark and f in kHz.

The formula for calculating the critical bandwidth for each frequency is [44, 17]:

$$BW_z(f) = 25 + 75 \cdot (1 + 1.4 \cdot f)^{0.69} \tag{6}$$

with bandwidths in Hz and frequencies in kHz.

[2] Although the original name of the scales are different, to pave the way for later notation, we are going to introduce our own names for the logarithmic scales, resp. m, z and ERB rate

2.2.3 ERB and ERB-rate

The ERB was defined in [33, 14] as a more adjusted measurement of the critical band:

$$BW_{ERB}(f) = 6.23 \cdot f^2 - 93.39 \cdot f + 28.52 \,(\text{f in kHz}) \tag{7}$$

Based upon these bands a new logarithmic scale may be defined, the *ERB-rate [33]*

$$F_{ERB}(f) = 11.17 \cdot \log \left| \tfrac{f+0.312}{f+14.675} \right| + 43.0 \,(\text{f in kHz}) \tag{8}$$

or the ERB_N number, [14, 34]:

$$ERB_N number(f) = 21.4 \log(4.37f + 1) \tag{9}$$

This scaling is at the base of the Gammatone filterbank, an alternative to the one employed by MFCC that will be employed in our tests (see section 6). This filterbank is defined in the time domain by its impulse response[35]:

$$f(t) = kt^{n-1} exp(-2\pi Bt) cos(2\pi f_c t + \phi) \tag{10}$$

where k defines the output gain, n is the order of the filter (in the range 3-5 the filter is a good approximation of the human auditory filter), B defines the bandwidth, f_c is the filter's central frequency and ϕ is the phase.

2.3 Masking

Masking is a phenomenon whereby the perception of some frequency is affected by another frequency, the *masker frequency* to the extent that masked frequencies may dissapear from perception. Our conceptual take is that all other masking behaviours (with narrow and wide band maskers) can be obtained by appropriated addition of tone maskers, so we will only consider this type of maskers herein. A masking tone will be defined as

$$t_{F_m}(F) = L_m \delta(F - F_m), \tag{11}$$

where F is a choice of a logarithmic transformation as the ones presented in section 2.2, L_m is the SPL of the tone and F_m is the appropriately scaled masker frequency.

Cochlear masking has been studied mainly as it regards the influence of some frequencies on others simultaneously present in the spectrum, or *simultaneous masking*, or as regards the influence of the same frequencies at different time instants, or *temporal masking.*

2.3.1 Simultaneous masking

Simultanous masking is the minimum sound pressure level of a *test sound, probe or signal* (normally a pure tone) that is audible in the presence of a *masker*. By varying the frequency of the probe throughout the spectrum a *masking pattern* may be obtained. As it happens, the shape and sound pressure level L_m of the masker is quite determinant of the masking pattern.

Regarding the change of masking with masker parameters, [43] note that simultaneous masking is better represented in logarithmic scales.

Indeed, Figure 1 shows the masked thresholds for masker tones of the same $L_m = 60dB\,SPL$ at different frequencies in linear (a), log-linear (b) and bark scale (c), (in variables f, $F = \log f$ and F_z, see section 2.2.2, respectively). We can see a greater regularity in spacing an masker slopes in the latter, a consistent finding in other auditorily-motivated scales.

Similarly, simultaneous masking produced by narrow-band maskers is level-dependent and therefore has clearly a non-linear effect, as shown in Figure 2.

2.3.2 Temporal masking

Premasking (or backwards masking) occurs before the appearance of the masker, while *postmasking (or forward masking)* happens *after* the masker. Premasking all but disappears around 50ms before the masker while the influence of postmasking may last as far as 500ms after the masking [11, p. 196]. Hence, premasking might be important for the masking of occlusives or affricates while postmasking will definitely be an issue for the longer-lasting sounds like vowels or nasals, perhaps even fricatives.

Figure 3 shows the conventions to describe temporal masking for long maskers. Δt with positive and negative values describes the temporal delay since masker onset time, while t_d describes time delay after masker turn-off.

2.3.3 Premasking

Premasking is noticeable about 20ms prior to the masker regardless of its level. This would call for a dependency on level (to make the growing exponential rate higher the higher the level), but we model it with a constant slope of 8ms to mimic postmasking with short bursts of duration less than 5ms (see below).

2.3.4 Postmasking

There is the question of modelling post-masking with different lengths of masker :

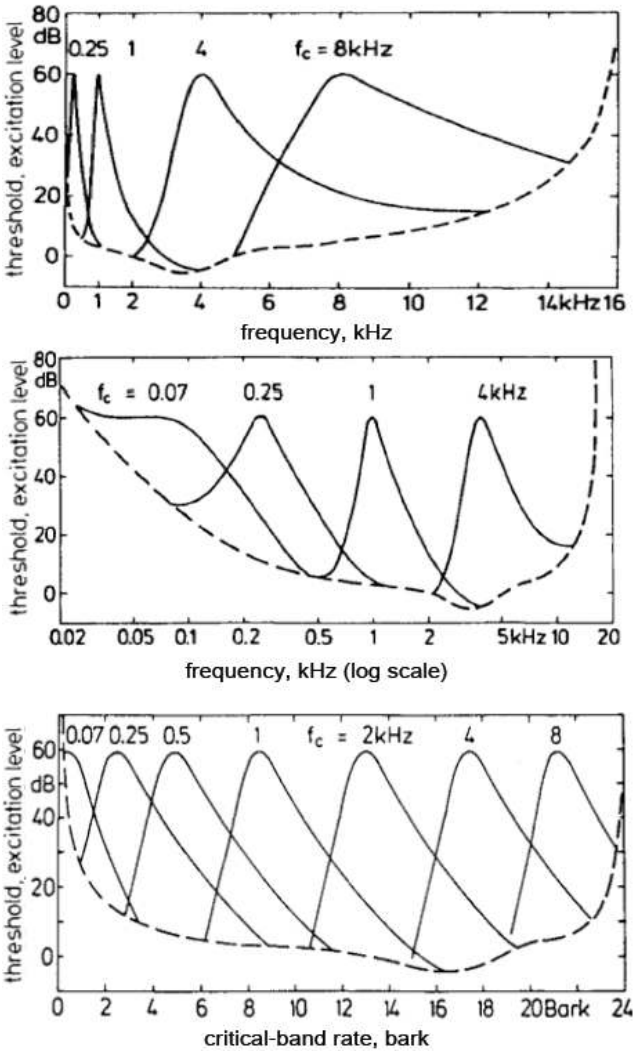


Fig. 1: Masking curves for a 60dB masker along the auditory range[taken from 45] in linear (a), log-linear (b) and bark scale (c).

- there is almost no decay for the first 5ms after the masker is switched off. The values amount to those observed for simultaneous masking [10, p. 83]. This would seem to rule out exponential decay as a model for postmasking.
- forward masking for a long duration masker (bigger than 5ms) lasts for 200ms regardless of noise level.
- For a uniform masking noise of 60dB, when the masker lasts for around 200ms the decaying time of the masker increases [10].

We are going to assume that the important model is that of masking evoked by short-duration maskers and let an additive model account for the decrease in slopes for longer masker durations.

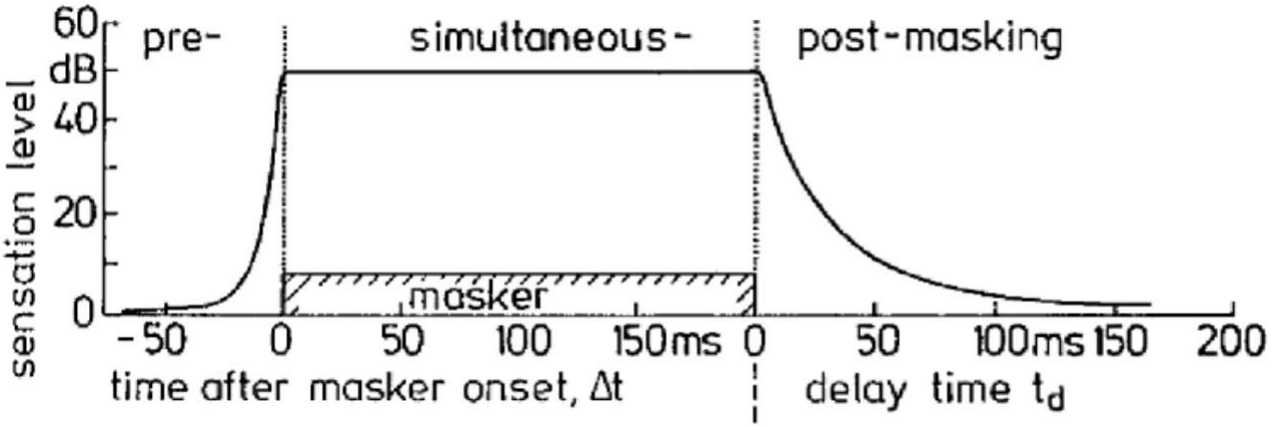


Fig. 3: Regions within which premasking, simultaneous masking and postmasking occur [from 10, fig. 4.17, p. 78].

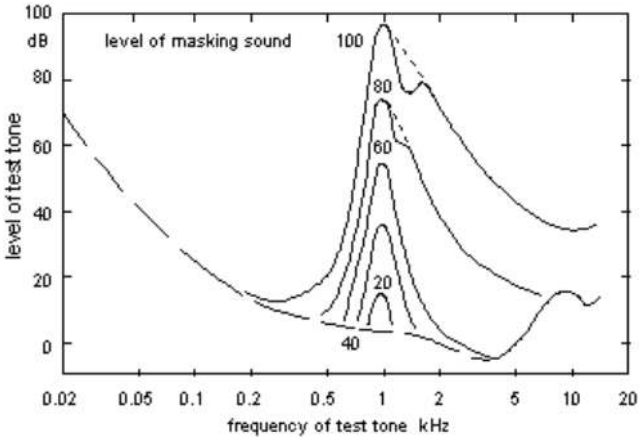


Fig. 2: Threshold in quiet and masking curves for a 1kHz masker [taken from 45].

A fitted model for post-masking is presented in [26] and later taken on by [17]:

$$M(t_d, L_m) = a(b - \log t_d)(L_m - c) \tag{12}$$

where M is the amount of masking, t_d is the signal delay in ms as depicted in figure 3, L_m is the masker level in dB SL, and a, b and c are parameters obtained by fitting the curve to the data. In particular,

- $a(L_m)$ is related to the slope of the time course of masking for a given masker level L_m.
- b is the logarithmic of the probe-masker delay intercept.
- c is the intercept when masker level is expressed in dB SL.

Note that this model disregards the variation of amount of masking with masker frequency.

This model has to be extended to take into consideration simultaneous masking.

2.3.5 Addition of the Threshold in Quiet

The actual masker threshold caused by any masker in the previous models is seen as linear in logarithmic delay and logarithmic frequency for non-simultaneous and simultaneous masking, respectively. This would entail very low masking amounts for delays (resp. frequencies) far away from the time-frequency of the masker.

Of course, this has to be corrected with the so-called *threshold in quiet (THQ)* to achieve the correct threshold surface. This threshold under which no sound is perceived can be modelled as [17],

$$\begin{aligned}\mathrm{THQ}_{T_S \geq 500}(f) &= 3.64 f^{-0.8} + 6.5 e^{(f-3.3)^2} + 0.0001 f \\ \mathrm{THQ}_{T_S < 500}(f,) &= \mathrm{THQ}_{T_S \geq 500}(f) \\ &\quad + (7.53 - 6.5 \cdot 10^{-3} f^3) \log_{10}(500 - T_s)\end{aligned}$$

where T_S is the duration time of the masker employed in the empirical determination of these expressions.

The combination that immediately comes to mind, the maximum:

$$L_M(t_d, f, L_m) = \max(M(t_d, L_m), THQ(f)) \tag{13}$$

is clearly wrong on experimental accounts: the skirts of the presented masker seem to taper off smoothly towards the threshold-in-quiet. Nonetheless, it is an acceptable approximation in most frequency ranges that we have assumed as a flooring in the spectrograms resulting from our processing.

2.3.6 Final considerations

It seems that the masking capabilities of the cochlea co-evolved in the presence of a noise that has the peculiarity of raising masking thresholds uniformly, that is giving a flat frequency response. This noise, adequately

named as *uniformly masking noise* (UMN) [42] has the aspect of a broadband lowpass noise (with cutoff frequency in the range of vocal frequencies). Interestingly, this can also be considered an idealised version of the spectrum of human speech.

For this reason and the preceding discussion, in this paper we will model the response of the cochlea to tone maskers and trust in morphological processing mechanism to provide for the flattening action on the masked thresholds, as will be explained in section 4.

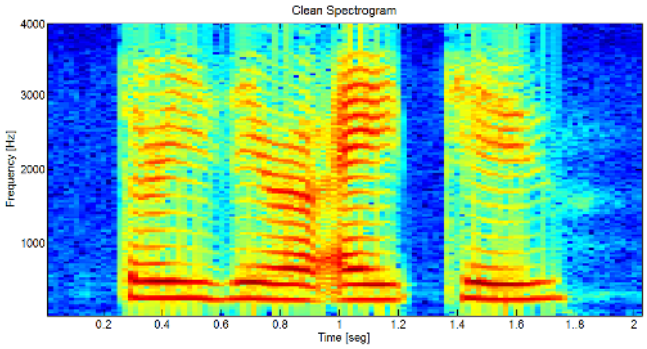
3 Spectrogram and Spectral Subtraction

As revealed in section 2 the effects of auditory masking can be observed both in time and frequency domains requiring a two dimensional representation to jointly consider the two of them. In this situation, a spectrogram is a natural choice since it expresses the speech signal spectral energy density as a function of time [28].

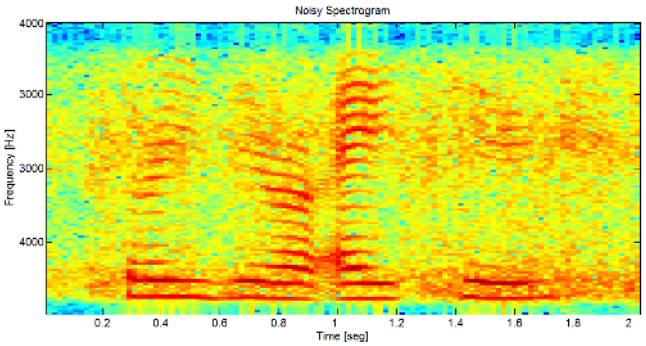
This popular depiction shows the temporal evolution of formant positions, harmonics and other components of speech. The spectrograms are usually displayed as gray-scale or heatmap images and typically, the larger energy magnitudes in the spectrum are displayed in white (or warm, in case of heatmaps) colours and the valleys (e.g. silences) in dark (or cold) colours. This is illustrated in Figures 4a and 4b by the spectrograms for clean and noisy signals.

On the other hand, the application of an auditory motivated scale in the frequency domain (see section 2.2) produces a more uniform representation of the simultaneous masking effects (see section 2.3.1) that is certainly more amenable for a computational modelling since the masking threshold becomes *almost* independent of the scaled frequency. An auditory spectrogram (sometimes referred as *cochleogram*) substitutes the usual linear spectral representation (resulting from a Discrete Fourier Transform) by auditory motivated filterbanks uniformly distributed in a scaled frequency.

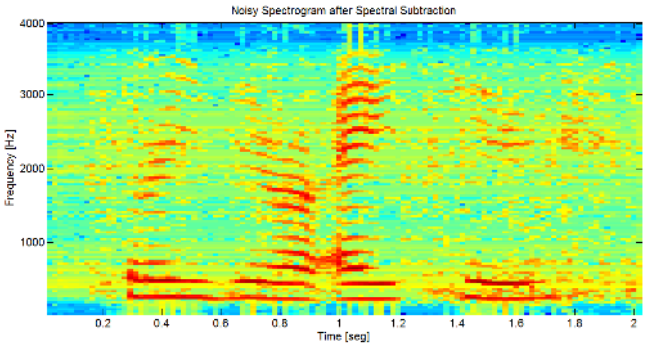
Another preprocessing technique employed in this paper is the conventional Spectral Substraction (SS) procedure [3] that applied on the *noisy spectrogram* as can be observed in figure 4c will be regarded as our baseline system. However, this method is known to exhibit what is called *musical* noise, i.e., rough transitions between the speech signal and the areas with removed noise become noticeable and unpleasant to a human listener. Among other aspects, our proposal aims at attenuating this pernicious behavior while preserving the main speech features.



(a) Clean Spectrogram



(b) Noisy Spectrogram



(c) Noisy Spectrogram + Spectral Subtraction

Fig. 4: Resulting spectrograms of a noisy utterance with added metro noise at 10dB SNR after spectral subtraction.

4 Morphological Filtering of Speech Signals

Mathematical Morphology is a theory for the analysis of spatial structures [39] whose main application domain is in Image Processing as a tool for thinning, pruning, structure enhancement, object marking, segmentation and noise filtering [15, 8].

It may be used on both black and white and gray-scale images, and in this paper, we report on its use for *morphological filtering (MF)*.

4.1 Morphological filtering for continuous signals

The basic operations in gray-scale morphology are *erosion* and *dilation*. Let $S(f,t)$ and $M(f,t)$ be two-dimensional signals. Their erosion $(S \ominus M)$ and dilation $(S \oplus M)$ are defined as the convolutions in $\mathbb{R} \cup \{\pm\infty\}$,

$$(S \ominus M)(f,t) = \bigwedge_{(\varphi,\tau)\in\mathbb{R}^2} \{S(f,t) - M(f-\varphi, t-\tau)\}$$
$$(S \oplus M)(f,t) = \bigvee_{(\varphi,\tau)\in\mathbb{R}^2} \{S(f,t) + M(f-\varphi, t-\tau)\}$$

where $\wedge$ is the min operation and $\vee$ the max operations extended adequately to operate on infinite values. Erosion is used to shrink or reduce objects, while dilation, being the dual to erosion, produces an enlarging. Both are irreversible.

Opening $S \circ M$ and *closing* $S \bullet M$ are just the two compositions generated by erosion and dilation with a fixed M, called the *structuring element (SE)*

$$S \circ M = (S \ominus M) \oplus M \qquad S \bullet M = (S \oplus M) \ominus M \ .$$

Note also that for fixed SE M opening (resp. closing) is an idempotent, decreasing (resp. increasing) operator, that is an *interior (resp. closure) operator*,

$$S \circ M \leq S \qquad S \bullet M \geq S$$
$$S \circ M = (S \circ M) \circ M \qquad S \bullet M = (S \bullet M) \bullet M \ .$$

Opening and closing are used to remove small objects in images, typically noise, their behaviour with respect to, for instance, salt and pepper noise, being dual to each other.

4.2 Morphological filtering for discrete signals

We use S with implicit frequency band index n and temporal frame k to represent an adequate discretization $S(n,k)$ of signal $S(f,t)$. Then erosion and dilation can be represented as matrix operations

$$S \ominus M = \{p \in \mathbb{R}^2 \mid p = m - s, m \in M,\ s \in S\}$$
$$S \oplus M = \{p \in \mathbb{R}^2 \mid p = s + m, s \in S, m \in M\}$$

Opening and closing adopt similar forms as matrices as they had for operators

$$S \circ M = (S \ominus M) \oplus M \qquad S \bullet M = (S \oplus M) \ominus M \ .$$

4.3 Discrete morphological spectro-temporal filtering of speech

Let S be a discretized spectrogram and M a specific discrete spectro-temporal masker function taken as structuring element. As in [9], we use the opening operator to try to remove the remaining noise and enhance any component of the speech signal to obtain an emphasized spectrogram that is subsequently added on the (possibly de-noised) spectrogram to produce the filtered speech signal,

$$\hat{S} = S + S \circ M \ .$$

Notice from an example such as that of Figure 5.(a) the irregular shapes of the *acoustic objects* of the spectrogram (i.e. formant and harmonic modulations).

We decided to test different SEs to try to capture such dynamics. For a first choice, the filtered spectrogram of Figure 5.(f) was obtained by pixel-wise adding those versions of the spectrogram obtained by morphological filtering with the three different SEs of Figure 6, at angles 0°, 45°and 90°, resulting in Figures 5.(b), (c), and (d), respectively. The emphasized spectrogram in Figure 5.(f) was then obtained by normalizing this filtered spectrogram and adding it to the original (de-noised) spectrogram.

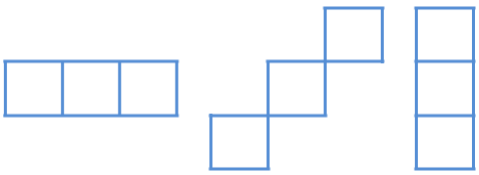
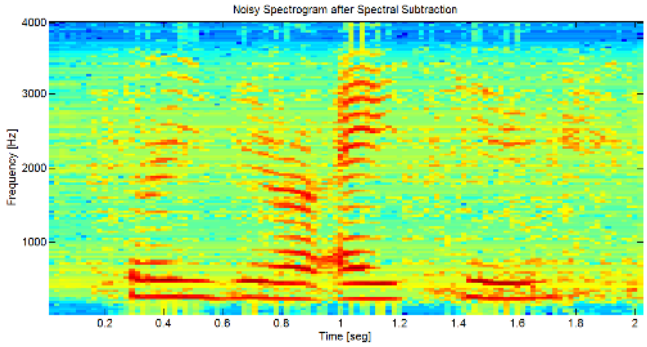


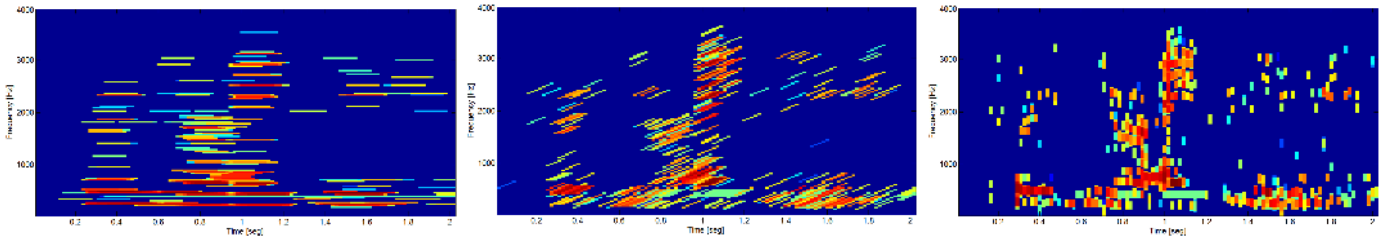
Fig. 6: Anisotropic SE: rectangles of different sizes and angles.

Our second choice of structuring element is the single flat anisotropic SE in Figure 7. Its design is inspired by the masking effects of the human auditory system (HAS) both in time and frequency (see §2, [45, 12]). On the auditory-inspired frequency axis, the spread of masking for simultaneous masking was modelled symmetricall, as in figure 1. On the time axis, however, the masking effect is asymmetrical (see figure 3) as per the experimental data of Section 2.3.2. Outside these well-explored phenomena, that is in places where a mixture of simultaneous- and temporal-masking would occur, the behaviour was extrapolated as shown in Figure 7.

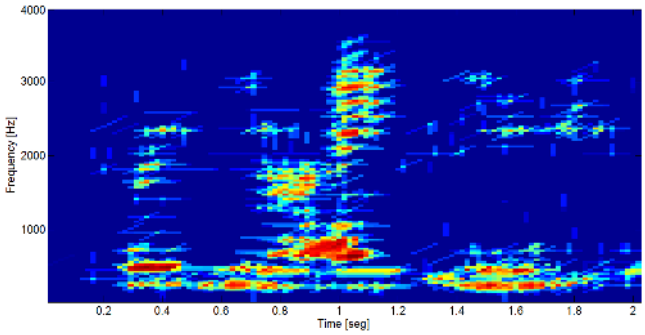
The procedure to obtain and combine the emphasized spectrogram with the spectrogram was the same already explained for the first SEs.



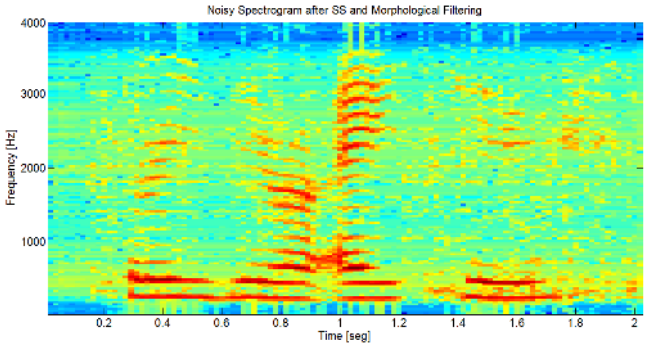
(a) Noisy Spectrogram + SS (Figure 4.4c)



(b) Horizontal Rectangle (c) Tilted Rectangle (d) Vertical Rectangle



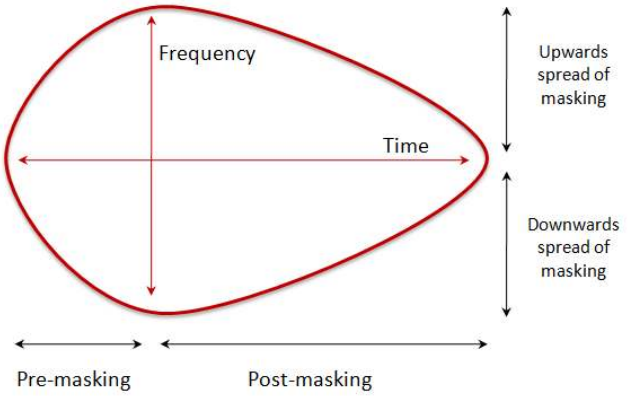
(e) The addition of (b), (c) and (d)



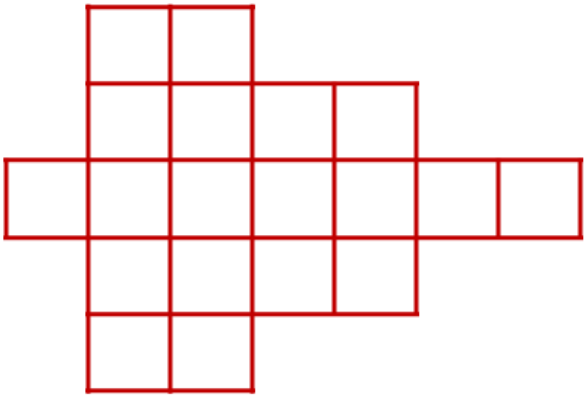
(f) The addition of (a) and (e)

Fig. 5: Morphological filtering on the noisy spectrogram of Figure 4(b), reproduced in (a) for comparison; (b), (c) and (d) are the result of filtering with the different anisotropic structural elements in Figure 6, while (e) is their pixel-wise addition and (f) the emphasis (addition) of (a) with (e).

The third choice was a single non-flat improvement on the second one: all the pixels were weighted as shown in Figure 8, but the rest of the procedure remained the same.



(a) Continuous masker-inspired structural element considering pre- and post-masking as well as the upwards and downwards spread of simultaneous masking imposed by a certain flooring



(b) Discretization of masker

Fig. 7: Obtaining an auditorily-motivated structural element: top, the estimated effect of a masker in the human auditory system; bottom, the flat structuring element designed to emulate that response

The first two anisotropic SEs of Figures 6 and 7 were used in the speech enhancement task of Section 5. But for the ASR task of Section 6 we only report on the performance of the HAS-motivated anisotropic SEs of Figures 7 and 8, reflecting the fact that they obtained better results in the Speech Enhancement task.

In all cases, values below a certain threshold in the resulting combined are replaced by a threshold value to avoid introducing a *musical* noise on the filtered signal. This models the existence of a thresholding mechanism as seen in Section 2.3.5.

5 Application to Speech Enhancement

In this section we present the evaluation of our proposed method, morphological filtering with different anisotropic SEs (Figs. 6 and 7), on a speech enhancement

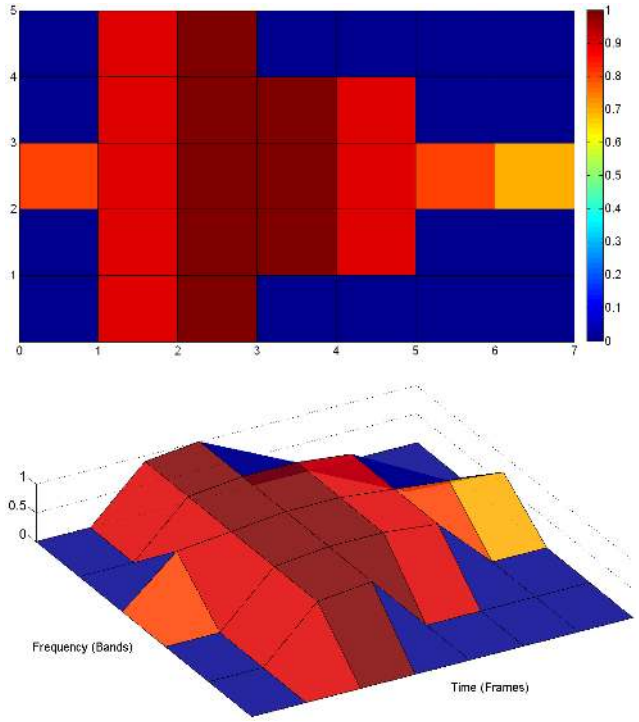


Fig. 8: Top panel shows the 2-D view of a structural element. Bottom panel shows the 3-D view. The color represent the weight of each pixel in the morphological operations.

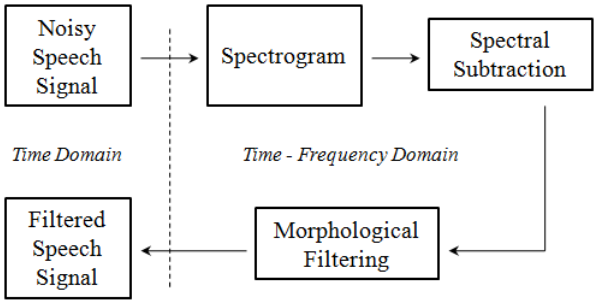


Fig. 9: Filtering proposed, step by step.

task. A block diagram of our proposed procedure can be observed in Figure 9.

5.1 General Description

The first step is to obtain the spectrogram of the noisy speech signal, sampled at 8 kHz. A resolution of 128 pixels (256-point FFT) is used on every spectrogram as it was empirically determined to be appropriate for this task. Next, a conventional SS is applied on the *noisy spectrogram* and the contrast of the resulting gray-scale image increased. The idea behind these operations is to emphasize the speech signal over the remaining noise to

make it easier for the subsequent morphological filtering process. This last is performed by applying an opening operation. Finally, the filtered signal in the time domain is recovered using a conventional overlap-add method.

5.2 AURORA Database

The evaluation of the filtered signals with the proposal method was conducted on the AURORA Project database [20] which makes use of a speech database based on TI digits with artificially added noise over a range of SNR's.

We have considered four types of noise: metro, car, airport and restaurant noise. We employed around a thousand speech files, individually contaminated with additive noise at 5 different values of SNR (-5dB, 0dB, 5dB, 10dB and 15dB). A clean speech signal with added noise (regardless of the SNR) is called in this paper *noisy signal*.

5.3 Estimation of Perceptual Quality with Objective Quality Measures

We used three objective quality measures (OQM) to evaluate the filtered signals:

- Signal Distortion (*Sig*), adequate for the prediction of the distortion on speech.
- Background Noise (*Bak*), for predicting background intrusiveness.
- Overall Effect (*Ovl*), for predicting the overall quality.

All the OQM ([21], [22]) are evaluated using a five point scale where 1 the worst scenario and 5 the best. The OQM consist of linear combinations of the following measures (Table 1):

- Perceptual Evaluation of Speech Quality (PESQ): recommended by the ITU-T for end-to-end speech quality assesment, the PESQ score is able to predict subjective quality with good correlation in a very wide range of conditions, that may include noise, filtering, coding distortions, errors, delay and variable delay [2], [38].
- Weighted-Slope Spectral Distance (WSS): computes the weighted difference between the spectral slopes in each frequency band. The spectral slopes are obtained as the difference between spectral magnitudes in dB [27].
- Log-Likelihood Ratio (LLR): Also referred to as the Itakura distance, is a measure of the perceptual difference between an original spectrum (in our context, the clean speech signal) and a modified version of that spectrum (the filtered speech signal) [37].
- Segmental Signal-to-Noise Ratio (segSNR): This frame-based measure is formed by averaging frame level SNR estimates [18].

Table 1: Weight of each measure on the *Sig*, *Bak* and *Ovl*.

Measures	PESQ	WSS	LLR	segSNR	Constant
Sig	0.603	-0.009	-1.029	0	3.093
Bak	0.478	-0.007	0	0.063	1.634
Ovl	0.805	-0.007	0.512	0	1.594

5.4 Experiments and Results

In order to evaluate the performance of our method we used the measures mentioned in section 5.3 with the code available in [29] and considering the clean speech signal as the reference. We have compared five different combinations. The first one corresponds to spectral subtraction (*SS*) and the other four correspond to different morphological filtering: black and white mask with a isotropic SE (*BW & iSE*), black and white mask with anisotropic-SE (*BW & aSE*), gray-scale mask with anisotropic-SE (*Gray & aSE*) and gray-scale mask with anisotropic-SE-2 (*Gray & aSE-2*). The last two are the proposed methods: anisotropic-SE are the rectangles and anisotropic-SE-2 is the HAS inspired by.

Results are presented using the following relative measure:

$$\Delta[\%] = 100 \cdot \frac{(FS - NS)}{NS} \tag{14}$$

in which *FS* is the filtered signal and *NS* is the noisy one. Positive increments imply an improvement and negative a signal degradation with respect to the noisy signal.

Overall, similar trends have been observed for all of the noises being the results of *car* and *metro* on the one side and *restaurant* and *airport* on the other, very similar. Therefore, we have chosen *metro* and *airport* as a representative sample of them.

5.4.1 Metro Noise

Results for Metro noise and several SNRs in terms of the relative measures with respect to the noisy signal are shown in Figure 10.

The method with gray-scale mask and anisotropic SE (*Gray & aSE*) provides the best performance for

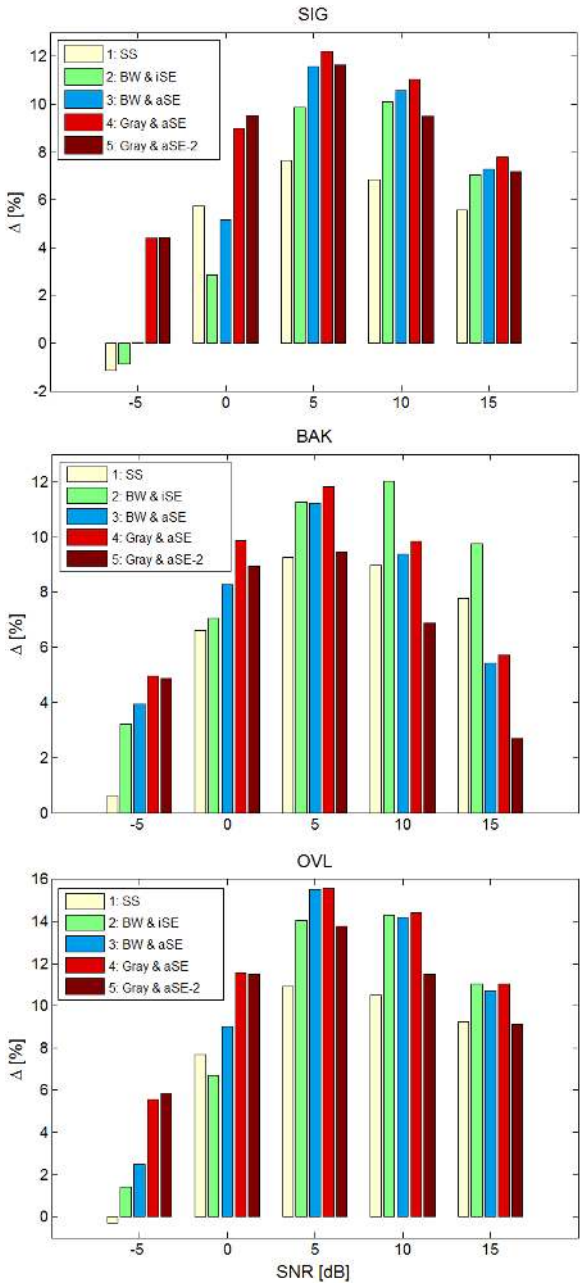


Fig. 10: From the top panel to the bottom panel: Relative measures for the Objective Quality Measures. Five different methods (SS: Spectral subtraction, BW: Black and white mask, Gray: Gray-scale mask, iSE: Isotropic SE, aSE: Anisotropic SE). Metro Noise was used.

high SNRs in terms of *Sig*. The method (*Gray & aSE-2*) is just better in low SNRs. The largest margin with respect the other 3 methods is obtained for SNR = −5dB.

With respect to the *Bak* measure, the *Gray & aSE* method achieves the best performance for SNRs of −5dB, 0dB and 5dB. However, the filtering with black and white mask and isotropic SE (*BW & iSE*) reaches the highest values of *Bak* for higher SNRs (10dB and 15dB). It is worth mentioning that this OQM factors in segmental SNR (see Table 1), which is known to be very sensitive to misalignments.

Best results for the *Ovl* measure are obtained for SNRs of 0dB, 5dB and 10dB when using (*Gray & aSE*) filtering. For SNR = −5dB (*Gray & aSE-2*) is the best. For SNR = 15dB the (*BW & iSE*) method is slightly better.

In summary, for the Metro noise, the proposed methods (and in general, the use of anisotropic structural elements) provides the best performance for low and medium SNRs (−5dB, 0dB and 5dB). For higher SNR where the speech signal may not need to be denoised, the filtering with black and white mask and isotropic SE presents a similar performance in comparison to other methods or slightly better, in terms of *Bak*.

5.4.2 Airport Noise

Figure 11 shows results for airport noise and several SNRs in terms of relative *Sig*, *Bak* and *Ovl* measures.

First of all, it is worth mentioning that for low SNRs, all the evaluated methods produce degradations in the quality of the processed signals. One possible explanation to this fact is the acoustic nature of the Airport environment in which babble noise is present. Spectrograms of the babble noise show the typical energy distribution of speech, making more difficult the denoising of the speech signals so contaminated.

As it can be observed, in terms of *Sig* and *Ovl* measures, (*Gray & aSE*) and (*Gray & aSE-2*) methods achieve the best performance, but the last one is more suitable for low range of SNRs. For the *Bak* measure, (*Gray & aSE-2*) and (*Gray & aSE*) methods provides the highest performance in low and medium SNRs (−5dB, 0dB, 5dB and 10dB) respectively (except of SNR = 15dB, in which both *SS* and (*BW & iSE*) filtering performs better).

6 Application to Automatic Speech Recognition

In this section we present the evaluation of our method in an automatic speech recognition task. In particular, we evaluate the performance of different spectro-temporal features derived from the morphological filtering of auditory-motivated spectrograms. Two different structuring elements are considered: the anisotropic SE *aSE-2* in Figure 7 and its weighted version, *aSE-3*, in Figure 8.

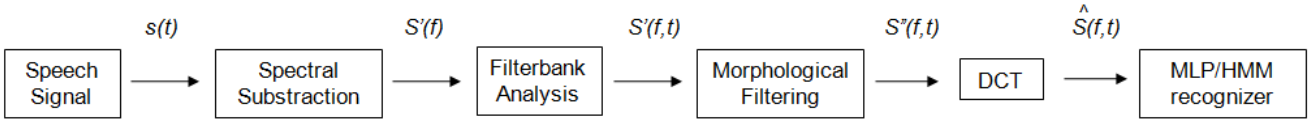


Fig. 12: Block diagram of the ASR system.

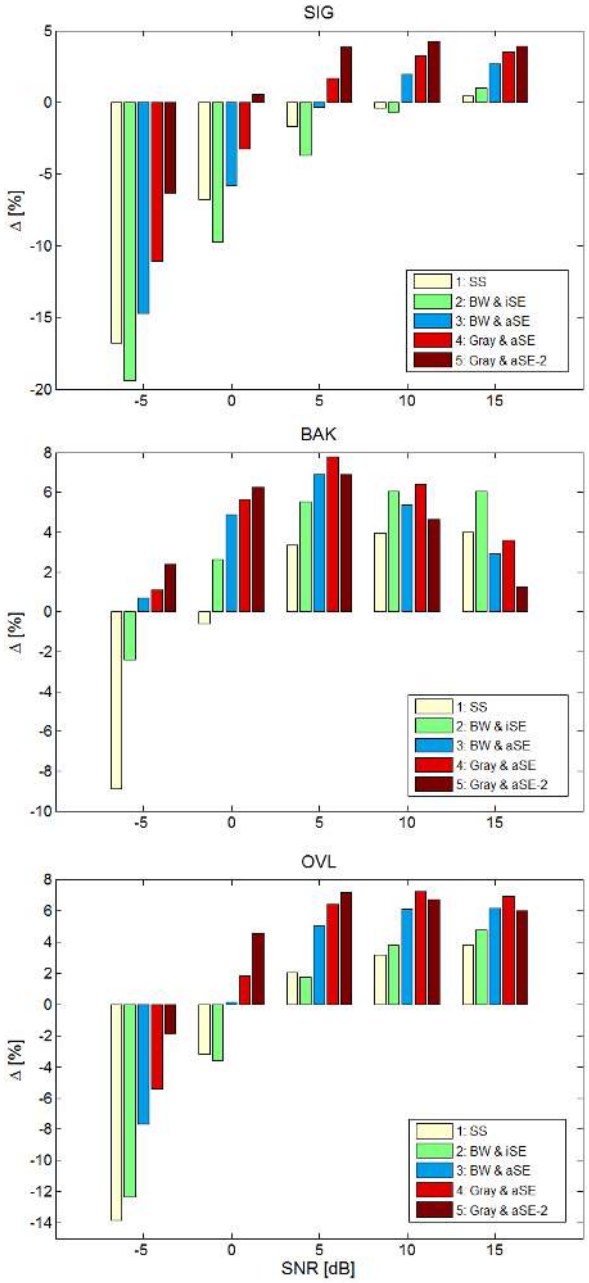


Fig. 11: From the top panel to the bottom panel: Relative measures for the Objective Quality Measures. Five different methods (SS: Spectral subtraction, BW: Black and white mask, Gray: Gray-scale mask, iSE: Isotropic SE, aSE: Anisotropic SE). Airport Noise was used.

6.1 General Description

The block diagram of the ASR system used in the experimentation is depicted in Figure 12. Firstly, a conventional Spectral Subtraction is applied on the noisy signal in order to emphasize as much as possible the speech signal over the remaining noise. Next, an auditory filtering is performed over this (partially) denoised spectrogram. Two different types of auditory filterbanks are considered: a set of triangular mel-scaled filters and a set of Gammatone filters (see subsections 2.2.1 and 2.2.3). In the next step, a morphological filtering using anisotropic elements is applied. Finally, in order to decorrelate the filterbank log-energies obtained in the previous stage, a Discrete Cosine Transform (DCT) is computed over them, yielding to Mel-Frequency Cepstrum Coefficients (MFCC) features in the case of using the mel-scaled filterbank and to Gammatone-based (GTC) features in the case of using the Gammatone filterbank.

For each type of features we train and test different MLP/HMM hybrid speech recognizers following the ISOLET testbed as described in subsection 6.3.

6.2 Feature Extraction

As mentioned before, two types of acoustic parameters are considered: MFCC and GTC features. In both cases, speech is analyzed using a frame length of 25 ms and a frame shift of 10 ms after pre-emphasis and Hamming windowing.

MFCCs and GTCs are computed from N-channels mel-scaled and gammatone filterbanks, respectively. After the DCT, we take the coefficients from C_0 to C_{12} plus the corresponding delta (Δ) and acceleration ($\Delta\Delta$) coefficients. Thus, both MFCC and GTC feature vectors are constituted by 39 components.

In both cases, we have performed experiments considering $N = \{40, 80, 128\}$ channels in the corresponding filterbanks.

6.3 ISOLET Testbed

In this application, we use the ISOLET testbed [13] and database [6]. ISOLET is a database of letters of the English alphabet spoken in isolation. The database consists of 7800 spoken letters (two productions of each letter pronounced by 150 different speakers). Specifically, we use a version called Noisy-ISOLET: the speech signals of ISOLET plus 8 different noise types at different SNRs (clean, 0dB, 5dB, 10dB, 15dB y 20dB). The noise types are: Speech babble, Factory floor noise 2, Car interior noise (volvo), Pink noise, F-16 cockpit noise, Destroyer operations room noise, Leopard military vehicle noise and Factory floor noise 1.

The experiments using the ISOLET testbed are performed over an hybrid MLP/HMM ASR system, whose fundamentals are described in [5]. A context of 5 frames is used (each one of 39 components), so the input of each MLP has 195 elements.

The hybrid MLP/HMM system is tested in two different conditions: in the first one, the system is trained using clean speech (*mismatched* case), whereas in the second one, the training set is composed of a balanced combination of speech contaminated with the different noises of the database at several SNRs (*matched* case). A 5-fold cross-correlation procedure has been employed to improve statistical significance.

6.4 Experiments with Morphological Filtering

This set of experiments was performed in order to study the impact of the Spectral Subtraction (SS), the Auditory filtering (either Mel-filterbank or Gammatone-filterbank) and the Morphological Filtering (MF) on the performance of the whole system. 40 bands have been employed in the auditory filterbank and the *aSE-2* anisotropic structuring element. Recognition results together with their corresponding 95% confidence intervals are shown in Table 2.

As can be observed, similar conclusions can be drawn from the results with MFCC and GTC features. First, in the *mismatched* condition, whereas SS clearly outperforms the corresponding baselines being the performance differences statistically significant, MF alone produces a slightly increment of the WERs. Nevertheless, the sequential use of both techniques (SS + MF) improves the performance of the system compared to the baseline and SS cases. In particular, the improvement achieved by SS + MF with respect to SS is statistically significant for the MFCC parameterization. For the *matched* condition, no significant improvements are achieved by using SS, MF or SS + MF and therefore our method seems to be more suitable for the *mismatched* case. It is worth pointing out that the *matched* case is harder to improve that the *mismatched* one, because higher speech recognition rates are achieved without any processing.

Table 2: Recognition results in terms of *WER* [%] The number of bands in these experiments was fixed to 40 (MFCC: Mel-frequency cepstral coefficients, GTC: Gammatone-based coefficients, SS: Spectral Subtraction, MF: Morphological Filtering).

Features	Mismatched	Matched
MFCC	51.80 ± 1.24	16.45 ± 0.92
MFCC + SS	40.85 ± 1.22	16.95 ± 0.93
MFCC + MF	54.03 ± 1.24	17.03 ± 0.93
MFCC + SS + MF	37.03 ± 1.20	17.05 ± 0.93
GTC	53.78 ± 1.24	17.15 ± 0.94
GTC + SS	40.28 ± 1.22	16.95 ± 0.93
GTC + MF	56.95 ± 1.23	17.43 ± 0.94
GTC + SS + MF	38.50 ± 1.21	16.85 ± 0.93

With respect to the auditory filterbank considered, MFCC achieves better results than GTC in all cases. However, the performance differences are not statistically significant.

Figure 13 shows the Recognition Rates achieved by the different techniques indicated in Table 2 as a function of the noise type for the *mismatched* condition. It can be observed that our method (bars 4 and 8) outperforms SS (bars 2 and 6) in all noise types for the MFCC parameterization and in 5 of the 7 noise types for the GTC parameterization. In this latter case, our method is only slightly worse for the babble and factory2 noises.

6.5 Experiments with different number of bands and structuring elements

As MF is applied on the output of the Auditory Filterbank analysis (in contrast to the speech enhancement task, in which MF is applied directly on the speech spectrogram), we carried out a set of experiments in order to analyse the influence of the number of frequency bands on the performance of the whole system. Note that there is a relationship between the number of bands and the size of the structuring element. Besides, we compared two structuring elements: the flat *aSE-2* and its nonflat version *aSE-3*.

Table 3 contains the corresponding *WERs* as well as the confidence intervals calculated for a confidence of 95%. For the *mismatched* condition, the behaviour of MFCC and GTC are different with respect to the number of bands. In GTC, the results slightly improve as

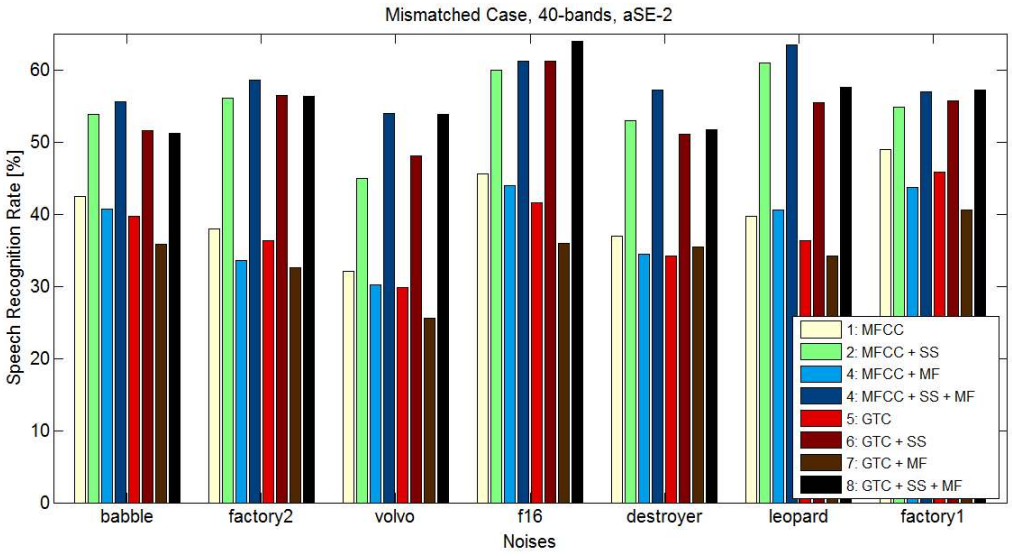


Fig. 13: Speech Recognition Rate [%] vs. Noise Types. Mismatched Case.

the number of bands increases for both structuring elements, although the differences are rather small. However, in MFCC, the best results are achieved with 128 bands and 80 bands, for *aSE-2* and *aSE-3*, respectively. This observation suggests that MFCC are more sensible to the size and shape of the structuring element.

In the other hand, for the *matched* condition, we obtain better recognition rates in all the cases with the nonflat SE. However, this condition seems to be sensitive to the number of bands. This result suggests the convenience of exploring new HAS-motivated nonflat SE taking into account the non-uniform masking effect explained in subsection 2.3.

7 Conclusions and future work

In this paper we have explored an alternative to the morphological filtering for speech enhancement and noise compensation proposed in [9] introducing elements of the HAS. In particular, we have proposed the use of morphological filtering with auditory-inspired anisotropic structural elements applied over gray-scale spectrograms. These ideas have been further extended to feature extraction for ASR.

On the one hand, for speech enhancement, results demonstrate that the proposed methods (using the anisotropic elements *aSE* and *aSE-2*) provide a better performance than the other alternatives for the SNR's of -5dB, 0dB and 5dB, a very important range of SNR's for speech enhancement. Besides, the proposed methods seem to be more suitable for non-stationary noise. However, subjective measures of the different alternatives could also shed more light into the evaluation procedure given that the objective estimates that we have employed in this paper have several limitations.

On the other hand, for automatic speech recognition, the combination of spectral subtraction and morphological filtering improves the recognition rates, regardless of the use of Mel or Gammatone filterbanks. Moreover, we think that there is more room for improvement in the morphological filtering stage. In particular, the incorporation of an automatic gain control (AGC) [30, 31] to Gammatone-filterbank features should be explored for ASR tasks. Other future lines of work include the use of other types of auditory filters, like those proposed in [23, 24] and alternative shapes for the anisotropic structural elements trying to more precisely emulate the masking effects (in time and frequency) of the human ear. The experimentation on real noisy signals instead of the artificially distorted ones employed in this paper is also desirable.

Acknowledgements This work has been partially supported by the Spanish Ministry of Science and Innovation CICYT Project No. TEC2008-06382/TEC.

References

1. Baker, J.: The Dragon system–an overview. IEEE Trans. on Acoustics, Speech, and Signal Processing **23**(1), 24–29 (1975)
2. Beerends, J., Hekstra, A., Rix, A., Hollier, M.: Perceptual evaluation of speech quality (PESQ) the new ITU standard for end-to-end speech quality assessment. Part II: Psychoacoustic model. Journal

Table 3: Recognition results in terms of *WER* [%]. All the experiments using Spectral Subtraction and Morphological Filtering with different number of bands and the Structuring Elements *aSE-2* and *aSE-3*.

Features	N-bands	Mismatched aSE-2	aSE-3	Matched aSE-2	aSE-3
MFCC	40	37.03 ± 1.20	37.02 ± 1.20	17.05 ± 0.93	16.53 ± 0.92
	80	37.70 ± 1.20	35.73 ± 1.19	16.93 ± 0.93	16.50 ± 0.92
	128	36.35 ± 1.19	38.60 ± 1.21	17.68 ± 0.95	17.00 ± 0.93
GTC	40	38.50 ± 1.21	38.38 ± 1.21	16.85 ± 0.93	16.70 ± 0.93
	80	38.13 ± 1.21	37.93 ± 1.20	17.53 ± 0.94	16.90 ± 0.93
	128	37.98 ± 1.20	37.40 ± 1.20	17.48 ± 0.94	16.90 ± 0.93

of the Audio Engineering Society **50**(10), 765–778 (2002)
3. Berouti, M., Schwartz, R., Makhoul, J.: Enhancement of speech corrupted by acoustic noise. In: Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'79., vol. 4, pp. 208–211. IEEE (1979)
4. ten Bosch, L., Kirchhoff, K.: Editorial note: Bridging the gap between human and automatic speech recognition. Speech Communication **49**(5), 331–335 (2007)
5. Bourlard, H., Morgan, N.: Hybrid HMM/ANN systems for speech recognition: Overview and new research directions. Adaptive Processing of Sequences and Data Structures pp. 389–417 (1998)
6. Cole R., M.Y., Fanty, M.: The isolet spoken letter database (2011). URL `http://cslu.cse.ogi.edu/corpora/isolet`
7. Davis, S., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. Acoustics, Speech and Signal Processing, IEEE Transactions on **28**(4), 357–366 (1980)
8. Dougherty, E.R., Lotufo, R.A.: Hands-on Morphologicla Image Processing, *Tutorial Texts in Optical Engineering*, vol. TT59. SPIE press (2003)
9. Evans, N., Mason, J., Roach, M., et al.: Noise compensation using spectrogram morphological filtering. Proc. 4th IASTED International Con. on Signal and Image Processing pp. 157–161 (2002)
10. Fastl, H., Zwicker, E.: Psycho-acoustics: Facts and Models, 3 edn. Springer (2007)
11. Florentine, M., Fastl, H., Buus, S.: Temporal integration in normal hearing, cochlear impairment, and impairment simulated by masking. The Journal of the Acoustical Society of America **84**(1), 195–203 (1988)
12. Flynn, R., Jones, E.: Combined speech enhancement and auditory modelling for robust distributed speech recognition. Speech Communication **50**(10), 797–809 (2008)
13. Gelbart, D., W., H., Holmberg, M., Morgan, N.: Noisy ISOLET and ISOLET testbeds (2011). URL `http://www.icsi.berkeley.edu/Speech/papers/eurospeech05-onset/isolet/`
14. Glasberg, B., Moore, B.: Derivation of auditory filter shapes from notched-noise data. Hearing Research **47**(1-2), 103–138 (1990)
15. Gonzalez, R., Woods, R.: Digital image processing (1993)
16. Greenberg, S.: The integration of phonetic knowledge in speech technology, *Text, Speech and Language Technology*, vol. 25, chap. From here to utility, pp. 107–132. Springer (2005)
17. Gunawan, T.S., Ambikairajah, E., Epps, J.: Perceptual speech enhancement exploiting temporal masking properties of human auditory system. Speech Communication **52**, 381—393 (2010)
18. Hansen, J., Pellom, B.: An effective quality evaluation protocol for speech enhancement algorithms. In: ICSLP, Sydney, Australia, pp. 2819–2822. Citeseer (1998)
19. Heckmann, M., Domont, X., Joublin, F., Goerick, C.: A hierarchical framework for spectro-temporal feature extraction. Speech Communication (53), 736–752 (2010)
20. Hirsch, H., Pearce, D.: The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW) (2000)
21. Hu, Y., Loizou, P.: Evaluation of objective measures for speech enhancement. In: Proc. Interspeech, pp. 1447–1450 (2006)
22. Hu, Y., Loizou, P.: Evaluation of objective quality measures for speech enhancement. Audio, Speech, and Language Processing, IEEE Transactions on **16**(1), 229–238 (2008)
23. Irino, T., Patterson, R.: A time-domain, level-dependent auditory filter: The gammachirp. Journal of the Acoustical Society of America **101**(1),

412–419 (1997)
24. Irino, T., Patterson, R.: A dynamic compressive gammachirp auditory filterbank. Audio, Speech, and Language Processing, IEEE Transactions on **14**(6), 2222–2232 (2006)
25. Jelinek, F., Bahl, L., Mercer, R.: Design of a linguistic statistical decoder for the recognition of continuous speech. IEEE Trans. on Information Theory **21**(3), 250–256 (1975)
26. Jesteadt, W., Bacon, S.P., Lehman, J.R.: Forward masking as a function of frequency, masker level, and signal delay. The Journal of the Acoustical Society of America **71**(4), 950–962 (1982)
27. Klatt, D.: Prediction of perceived phonetic distance from critical-band spectra: a first step. In: Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82., vol. 7, pp. 1278–1281. IEEE (1982)
28. Loizou, P.: Speech enhancement: Theory and practice. CRC-Press (2007)
29. Loizou, P.: Matlab software (2011). URL `http://www.utdallas.edu/~loizou/speech/software.htm`
30. Meddis, R.: Simulation of mechanical to neural transduction in the auditory receptor. J. Acoust. Soc. Am **79**(3), 702–711 (1986)
31. Meddis, R.: Simulation of auditory-neural transduction: Further studies. J. Acoust. Soc. Am **83**(3), 1056–1063 (1988)
32. Meyer, B., Kollmeier, B.: Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition. Speech Communication (53), 753–767 (2010)
33. Moore, B., Glasberg, B.: Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. The Journal of the Acoustical Society of America **74**, 750 (1983)
34. Moore, B., Glasberg, B.: A revised model of loudness perception applied to cochlear hearing loss. Hearing Research **188**(1-2), 70–88 (2004)
35. Patterson, R., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., Allerhand, M.: Complex sounds and auditory images. Auditory physiology and perception **83**, 429–446 (1992)
36. Peláez-Moreno, C., García-Moral, A., Valverde-Albacete, F.: Analyzing phonetic confusions using formal concept analysis. The Journal of the Acoustical Society of America **128**(3), 1377–1390 (2010)
37. Quackenbush, S., Barnwell, T., Clements, M.: Objective measures of speech quality. Prentice Hall Englewood Cliffs, NJ (1988)
38. Rix, A., Hollier, M., Hekstra, A., Beerends, J.: Perceptual evaluation of speech quality (PESQ), the new ITU standard for end-to-end speech quality assessment. Part I: Time-delay compensation. Journal of the Audio Engineering Society **50**(10), 755–764 (2002)
39. Serra, J., Soille, P. (eds.): Mathematical Morphology and its Application to Image Processing. Computational Imaging and Vision. Kluwer Academic (1994)
40. Stevens, S.S., Volkmann, J., Newman, E.B.: A scale for the measurement of the psychological magnitude of pitch. J. Acoust. Soc. Am. **8**, 185–190 (1937)
41. Yin, H., Hohmann, V., Nadeu, C.: Acoustic features for speech recognition based on gammatone filterbank and instantaneous frequency. Speech Communication (53), 707–715 (2010)
42. Zwicker, E., Feldtkeller, R.: The ear as a communication receiver. Acoustical Society of America (1999)
43. Zwicker, E., Jaroszewski, A.: Inverse frequency dependence of simultaneous tone-on-tone masking patterns at low levels. The Journal of the Acoustical Society of America **71**(6), 1508–1512 (1982)
44. Zwicker, E., Terhardt, E.: Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. The Journal of the Acoustical Society of America **68**, 1523 (1980)
45. Zwicker, E., Zwicker, U.: Audio engineering and psychoacoustics: Matching signals to the final receiver, the human auditory system. J. Audio Eng. Soc **39**(3), 115–126 (1991)