Morphologically Filtered Power-Normalized Cochleograms as Robust, Biologically Inspired Features for ASR

Fernando de-la-Calle-Silos, *Student Member, IEEE*, Francisco J. Valverde-Albacete, *Member, IEEE*, Ascensión Gallardo-Antolín, and Carmen Peláez-Moreno, *Member, IEEE*

***Abstract*—In this paper, we present advances in the modeling of the masking behavior of the human auditory system (HAS) to enhance the robustness of the feature extraction stage in automatic speech recognition (ASR). The solution adopted is based on a nonlinear filtering of a spectro-temporal representation applied simultaneously to both frequency and time domains—as if it were an image—using *mathematical morphology* operations. A particularly important component of this architecture is the so-called *structuring element* (SE) that in the present contribution is designed as a single three-dimensional pattern using physiological facts, in such a way that closely resembles the masking phenomena taking place in the cochlea. A proper choice of *spectro-temporal representation* lends validity to the model throughout the whole frequency spectrum and intensity spans assuming the variability of the masking properties of the HAS in these two domains. The best results were achieved with the representation introduced as part of the power normalized cepstral coefficients (PNCC) together with a spectral subtraction step. This method has been tested on Aurora 2, Wall Street Journal and ISOLET databases including both classical hidden Markov model (HMM) and hybrid artificial neural networks (ANN)-HMM back-ends. In these, the proposed front-end analysis provides substantial and significant improvements compared to baseline techniques: up to 39.5% relative improvement compared to MFCC, and 18.7% compared to PNCC in the Aurora 2 database.**

***Index Terms*—Auditory-based features, automatic speech recognition (ASR), cochlear masking models, morphological filtering, power normalized cepstral coefficients (PNCC), spectro-temporal processing.**

I. Introduction

THE remarkable ability of humans in speech recognition tasks under noisy conditions is still far above that of machines. In this context, several researchers have proposed that modeling the Human Auditory System (HAS) may be an adequate strategy to reduce the gap in performance.

It is well established that feature extraction methods for ASR need to take into account properties of the HAS to a certain extent: the well-known Mel-Frequency Cepstral Coefficients (MFCC) [1], for example, result from non-linear transformations of the frequency domain that mimic Fletcher's psychophysical transfer function [2], and include a triangular filterbank that emulates critical bands in the cochlea. Some other aspects, like the non-linear perception of sound intensity, are also incorporated by means of a logarithmic transformation of the spectrum.

Also widespread, Perceptually-based Linear Prediction (PLP) [3] is a pragmatic approach to model the auditory periphery that includes: resampling for frequency warping, bark-scale filter-bank, limited frequency resolution, pre-emphasis according to the threshold of hearing, amplitude compression and smoothing using linear prediction. The computational complexity of PLP feature extraction is similar to MFCC and sometimes provides better recognition accuracy.

There are plenty of other feature extraction methods that take into account the HAS, such as zero crossing peak amplitude (ZCPA) [4], average localized synchrony detection (ALSD) [5], perceptual minimum variance distortionless response (PMVDR) [6], invariant-integration features (IIF) [7], amplitude modulation spectrogram [8], sparse auditory reproducing kernel (SPARK) [9] or the well-known RelAtive SpecTrAl processing (RASTA) [10] that exploits the insensibility of human hearing to slowing varying stimuli by modeling the trend of the auditory periphery to emphasize the transient portions of incoming signals.

On the other hand, a number of detailed physiological models were proposed in the 1980s such as Seneff's auditory model [11] that mimics the nominal auditory-nerve frequency by employing 40 recursive linear filters implemented in cascade and also models the nonlinear transduction from the motion of the basilar membrane to the mean rate of auditory-nerve spike discharges, Ghitza's Ensemble Interval Histogram (EIH) model [12] uses the peripheral auditory model proposed by Allen [13] to describe the transformation of sound pressure into the neural rate of firing and focused on the mechanism used to interpret the neural firing rates, or Lyon's model [14], [15] where nonlinear compression, lateral suppression, temporal effects and correlograms are included.

Although these models do not generally provide improved performance on clean speech, they obtain better results than

Manuscript received January 13, 2015; revised May 22, 2015; accepted July 20, 2015. Date of publication August 05, 2015; date of current version August 28, 2015. This work was supported by an Airbus Defense and Space Grant (Open Innovation - SAVIER) and Spanish Government-CICYT projects TEC2014-53390-P and TEC2014-61729-EXP. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Bin Ma.

F. de-la-Calle-Silos, A. Gallardo-Antolín, and C. Peláez-Moreno are with the Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Leganés, 28912 Madrid, Spain (e-mail: fsilos@tsc.uc3m.es; gallardo@tsc.uc3m.es; carmen@tsc.uc3m.es).

F. J. Valverde-Albacete is with the Departamento de Leguajes y Sistemas Informticos, Universidad Nacional de Educación a Distancia, 28040 Madrid, Spain.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2015.2464691

conventional feature extraction methods when speech is degraded, for example, with added noise or reverberation. However, a usually higher computational cost and complexity (with a large number of parameters to be tuned) have prevented a more widespread adoption.

PNCCs [16], [17] have been proposed as an alternative to capture the essentials of the HAS without the complexity of full psychoacoustical models. They include the use of a power-law non-linearity that replaces the traditional logarithmic non-linearity used in MFCC coefficients and provides a better fit to the onset portion of the rate-intensity curve developed by the model of [18], a noise-suppression algorithm based on asymmetric filtering that suppresses background excitation and a module that carries out a *temporal masking* by placing a peak for each frequency channel and suppressing the instantaneous power if it falls below that of the envelope. As explained later in the paper, these features provide dramatic performance improvements over conventional MFCCs and their spectro-temporal representation—or *cochleogram*—will be the base for our developments.

Other methods also include procedures that emulate HAS masking: in contrast with PNCC that only includes temporal masking, *simultaneous or frequency masking* is considered in [19] where a frequency-dependent masking threshold is computed, or [20] that performs an estimation of the clean signal taking into account masking effects. In [21] both temporal and simultaneous masking are incorporated performing a time-frequency noise spectral subtraction.

Though most of the algorithms described above include spectro-temporal notions, these are incorporated in separate stages of the processing pipeline. The idea of simultaneously performing temporal and spectral analysis to yield so-called *spectro-temporal features* has lately emerged, e.g. spectro-temporal Gabor features [22], [23], [24], HIerarchical Spectro-Temporal (HIST) [25], spectro-temporal derivative features [26] or sparse spectro-temporal features [27]. Auditory-inspired representations in these domains are reviewed in [28].

Finally, noise robustness techniques are pervasive in ASR, some of them based on the (partial) suppression of background noise from the speech signal in a preprocessing stage. Most of these methods operate on the frequency-domain—like the already mentioned SS [29], Wiener filtering [30] or the minimum mean-square error short-time spectral amplitude estimator [31]—and attempt to enhance the speech signal without extensive modeling of the HAS properties.

As in these previously mentioned works, we also hypothesize that mimicking the Human Auditory System (HAS) may contribute to improve the performance of ASR systems in noisy conditions. Specifically, in this paper we model the masking behavior of the HAS to enhance the robustness of the feature extraction stage in ASR. Despite ingrained intuitions that masking deteriorates signal quality, we propound that it smooths away some noise and artifacts.

The three cornerstones of our procedure are first, the use of *mathematical morphology* operations to emulate the masking processing of the cochlea, second, the design of *a single auditory-inspired three-dimensional mask* independent of frequency and intensity and third, the use of an adequate *underlying spectro-temporal representation* of speech such that the non-linearities in frequency and intensity observed in the auditory masking phenomena are significantly equalized licensing a biologically meaningful application of the two previously mentioned elements.

In particular, our model filters a cochleogram—a spectro-temporal representation of speech—as if it were an image, allowing for the simultaneous processing of both dimensions, time and frequency. The morphological filtering procedure we propose aims to reproduce the masking properties of the HAS. For that purpose, the *mask*—or in mathematical morphology terminology, the *structuring element* (SE)—reproduces the spectro-temporal masking behavior as induced from well-known empirical measurements. Thus, the design of the SE is the crux of our approach.

Note that these empirical measurements were either carried out in the spectral or the temporal domains separately, but we need to extrapolate this to both dimensions. In this paper, we present various structuring element designs that aim at closely resembling the auditory masking phenomena taking place in the cochlea and we also refine our hypothesis that morphological filtering produces a smoothing of the spectro-temporal envelope that better models the masking behavior of the cochlea.

In [32], [33] we presented some evidence of this using Morphological Filtering of speech spectrograms with a roughly-approximated SE. Such rough modeling already yielded an enhancement of the filtered speech both in terms of objective quality measures and ASR performance. Note that, although some work has been carried out in the field of morphological processing of speech spectrograms using dilation across spectral lines to reduce spectral fluctuations [34], such efforts did not take into account the properties of the HAS.

Finally and for simplicity's sake, we employ a single mask across all frequencies and intensities despite the fact that the masking properties are frequency- and sound intensity-dependent [35], relying on the underlying spectro-temporal representation to accommodate these effects. The proper choice of this representation is essential in our feature extraction method. We have selected the one recently proposed in [16], [17] as part of the Power-Normalized Cepstral Coefficients (PNCC) in combination with conventional Spectral Subtraction (SS).

In summary, our contribution in this work is the simultaneous spectro-temporal emulation of the HAS masking phenomena by Morphological Filtering (MF) operations maintaining a low computational cost and complexity with very few tuning parameters. A key aspect is the design of a single bio-inspired three-dimensional SE that is used across the board unlike other spectro-temporal techniques that need larger numbers of different bases as in [22], [23], [24], for example, where a reduced set of temporal, spectral and spectro-temporal filters need to be chosen to make it feasible. For this single SE to remain invariant in frequency and intensity we rely on an underlying spectro-temporal representation that already accounts for that variability. In particular, we have borrowed that of PNCC—even improving the temporal masking there included—while maintaining a low computational complexity with respect to the PNCC baseline.

Regarding our previous works, on the one hand, the highly promising results on the Aurora 2 database noisy continuous digits task presented in [36] are now illustrated with a greater detail and, on the other hand, the performance of the proposed front-end on other different tasks, such as the Wall Street Journal and ISOLET databases, is also shown. The use of both conventional Hidden Markov Model (HMM) systems in the first two databases and an Artificial Neural Network/Hidden Markov Model (ANN/HMM) hybrid one on the third database, underlies the remarkable improvements of this feature extraction method across different domains and back-ends.

The rest of this paper is organized as follows: Section II introduces the theory and modeling alternatives for the underlying spectro-temporal representation, Section III explains the theoretical and empirical basis of cochlear masking, Section IV describes our three-dimensional model of this phenomenon introducing the basic terminology of Mathematical Morphology and the design of our biologically inspired SE. Finally Section V presents the results obtained in various datasets followed by some conclusions and further lines of research in Section VI.

II. Spectro-temporal Representation

As highlighted before, the underlying spectro-temporal representation—the cochleogram—where the morphological filtering will be applied needs to adopt the necessary frequency scaling and intensity normalization to allow for a single SE to be valid across the full spectrum and intensity range. Since our models have been tested in two different auditory-motivated frequency-scaled cochleograms $S(f,t)$ known as *Mel-frequency and Power Normalized based spectro-temporal representations*, respectively, next follows a detailed review of the possible alternatives.

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 realization stems from experiments to detect critical bands, the frequency bandwidth around a center frequency whose components affect the sound level and pitch perception of the center frequency.

In this light, the notion of an auditory filter-bank 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 bandwidths 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, and we will consider several scales of logarithmic frequency: *Bark*, *Mel* 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.

1) Critical Band and Critical-band Rate Scale: The Bark scale was first defined by [37]:

$$F_z(f) = \frac{26.8}{1 + \frac{1960}{f}} - 0.53, \tag{1}$$

where F_z is in bark units and f in Hz. The cochlear masking models described in Section III, which are derived from a set of psychoacoustic experiments, are defined in terms of the Bark scale.

2) The MEL Scale: The Mel scale [38] is a very well-known logarithmic transformation of the frequency scale:

$$F_m(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right), \tag{2}$$

where F_m is in mel units and f in Hz. This frequency transformation is in the core of the most popular ASR feature extraction procedure, the 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 explained in Section V.

3) ERB and ERB-rate: The ERB was defined in [39], [40] as a more adjusted measurement of the critical band:

$$BW_{ERB}(f) = 6.23 \cdot f^2 + 93.39 \cdot f + 28.52 (f \text{ in kHz}). \tag{3}$$

Based upon these bands a new logarithmic scale may be defined, the *ERB-rate* [41]

$$F_{ERB}(f) = 11.17 \cdot \log \left| \frac{f + 0.312}{f + 14.675} \right| + 43.0 (f \text{ in kHz}), \tag{4}$$

or the ERB number, [39], [40]:

$$ERB_N(f) = 21.4 \log(4.37 f + 1). \tag{5}$$

Alternatively, a filterbank can also be defined in the time domain by its impulse response, e.g. [42]:

$$h_{f_c}(t) = k t^{n-1} exp(-2\pi B t) cos(2\pi f_c t + \phi), \tag{6}$$

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.

This scaling is at the base of the Gammatone filter-bank used in PNCC, among others, an alternative to the one employed in MFCC that we will compare in our experiments (see Section V). According to [43], the impulse response of the Gammatone function provides an excellent fit to the human auditory filter shapes allowing a better modeling of the masking phenomena. Besides, PNCC [16] incorporate a medium-duration power bias subtraction and a power function nonlinearity to obtain the cochleogram, $S(f,t)$.

Dashed boxes in Fig. 1 contain the block diagrams of the two spectro-temporal representations considered in this work: Mel-frequency (left) and Power Normalized (right). The outputs of both submodules are the corresponding cochleograms $S(f,t)$ on which further processing with morphological filters is applied as explained in Section IV-C. Note that spectral subtraction (shadow block after STFT) is not part of the original

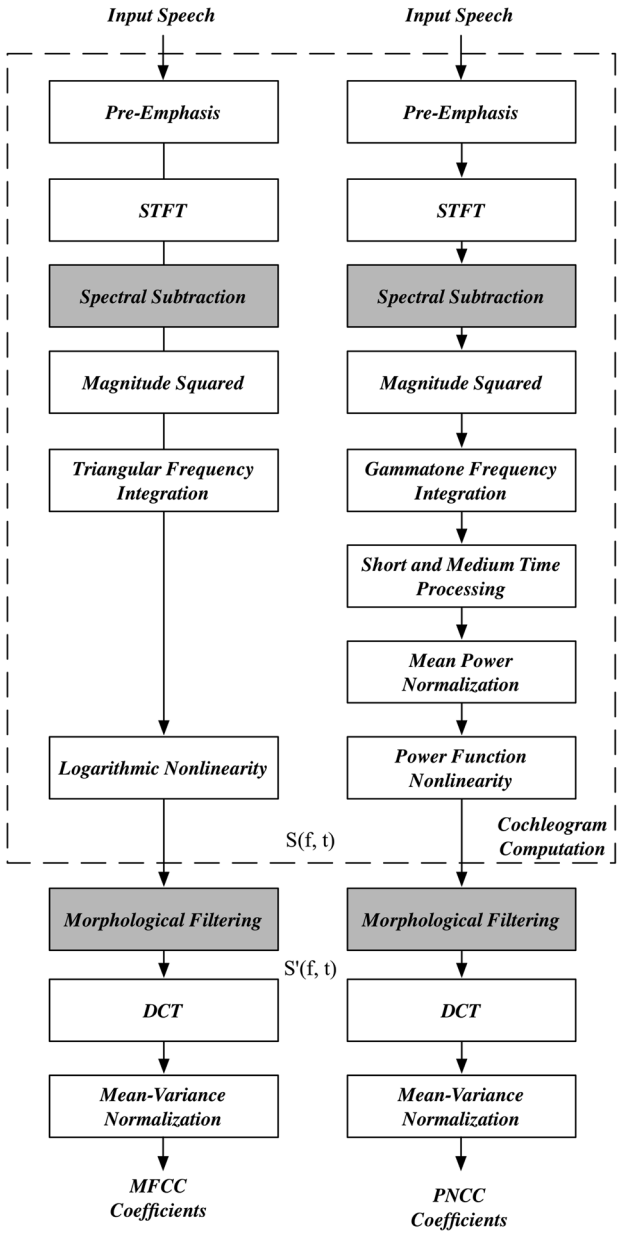

Fig. 1. Structure of the proposed front-end algorithm for the two spectro-temporal representations; the dashed boxes contain the submodules corresponding to the mel-frequency (left) and power-normalized (right) representations; the shaded blocks (Spectral Subtraction (SS) and Morphological Filtering (MF)) indicate the differences regarding conventional MFCC-based and PNCC-based feature extraction.

mel-frequency and power-normalized representations computations, but it is included here as a basic denoising technique (see Section IV-C).

III. Cochlear Masking Empirical Results and Models

The cochlea is the organ that converts the mechanical vibrations in the middle ear to neural impulses. The basilar membrane—the sensing structure that runs the length of the cochlea—has a particular frequency and time response [44].

Cochlear masking is the phenomenon whereby the perception of some frequency at a particular time instant, the *masked frequency*, is affected by the sound level of another, the *masker frequency*—possibly at a different time instant—, to the extent that masked frequencies may disappear from perception.

A masking tone will be defined as

$$s(F,t) = L_m \delta(F - F_m, t - T_m), \tag{7}$$

where F is in any of the transformed frequency scales introduced in the previous section, L_m is the sound pressure level of the tone, F_m and T_m are the masker frequency and time instant and δ represents the Dirac delta function.

Cochlear masking has been studied as the effect of a masker on simultaneously masked frequencies, *simultaneous masking*, and as the phenomenon whereby a masker affects non-simultaneous frequencies, *temporal masking*. Classical masking experiments concentrated in determining the amount of masking in either of these directions—frequency or time—in isolation. But it is important to notice that a given (masked) frequency is *always* being masked by maskers at different time instants—both from earlier and later maskers—and frequencies—both from lower and higher frequency maskers.

A. Simultaneous Masking

Simultaneous masking is defined as 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. An experimental fact is that 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, [45] noticed that simultaneous masking is better represented in logarithmic scales where the spacing and the masker frequency slopes extend more regularly to either side of the spectrum.

A simultaneous masking model can be extracted from Fig. 6.14 of [35] by fitting slopes for $L_m = 60$ dB in the Bark scale. We assume a constant L_m across all frequencies and intensities, relying on the underlying spectro-temporal representation to accommodate the frequency-intensity dependency of the masking properties.

B. Temporal Masking

Temporal masking has methodologically been treated as two separate processes: *premasking* occurs before the appearance of the masker while *postmasking* manifests itself after the masker is no longer present. It is well agreed-upon that premasking is noticeable about 20 ms prior to the masker, while the duration of postmasking extends well beyond 200 ms, perhaps as far as 500 ms [32].

Thus, premasking can be modeled as a constant slope of $+25$ dB/ms, starting 20 ms before the masker. Postmasking can be modeled with the fitted model for single masker-induced postmasking presented in [46],

$$M\left(t - T_m, L_m\right) = a(b - \log\left(t - T_m\right))(L_m - c) \tag{8}$$

where M is the amount of masking, t is measured in ms, L_m is the masker level in dB SPL, and a, b and c are parameters obtained by fitting the curve to the data. In particular,

- a is related to the slope of the time course of masking.
- b is the logarithmic of the probe-masker delay intercept.
- c is the intercept when masker level is expressed in dBL.

C. Smoothed Masking Responses

As suggested by the previous sections, an idealized masking model for a masker at (F_m, T_m) could be a cone with the appropriate decays in the (logarithmically-)scaled frequency and time coordinates. But findings consistently suggest a masking model that is smooth around (F_m, T_m), with sublinear decays close to this point and superlinear decays further away [35]: a sort of apex-smoothed cone.

At this point, it is worth mentioning that 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 [35].

We hypothesize that at the level of granularity at which the cochlear response is being observed this phenomenon is also present, and the masking response of a particular tone (F_m, T_m) must be the non-linear aggregation of many masking responses of other neighboring masking tones $(F_m + \Delta F, T_m + \Delta T)$ with $\Delta T \ll T_m, \Delta F \ll F_m$ which account for the smooth sublinear decay. This would manifest as a smoothness constraint for models of the masking response in the neighborhood of (F_m, T_m). This will be used in Section IV-B to constraint the SE.

IV. A Three-Dimensional Model of Cochlear Masking

A. An Overview of Morphological Processing

Mathematical Morphology is a theory for the analysis of spatial structures [47] whose main application domain is in Image Processing as a tool for thinning, pruning, structure enhancement, object marking, segmentation and noise filtering [48]. It may be used on both binary and grey-scale images.

To perform MF operations, we first convolve the image with a structuring element and then select the output value depending on the thresholded result of the convolution. In this paper, we apply MF on *cochleograms*, our underlying spectro-temporal representation, that will be processed as if they were images. This spectro-temporal representation is explained on Section II.

With the proper choice of SE, morphological operations on the cochleogram reproduce the phenomenon of auditory masking where the most prominent or salient elements of the cochleogram mask their surroundings in both the temporal and frequency domain.

Erosion and *dilation* are the basic morphological operations. Erosion is used to reduce objects, while dilation produces an enlargement and fills in small holes. Let S be the underlying spectro-temporal representation and M the three-dimensional structuring element, erosion is defined as: $S \ominus M$ and dilation: $S \oplus M$.

Erosion and dilation with a general structuring element require relatively simple algorithms and there are fast implementations that allow us to perform such operations efficiently. For gray-scale images, erosion is the minimum over the structuring element and dilation the maximum, respectively. For a pixel at (n, k) where n is the frequency bin and k the time step these operations can be defined as follows:

$$(S \ominus M)(n,k) = \min_{(\phi,\tau)\in\mathbb{R}^2} \{S(n,k) - M(n-\phi, k-\tau)\}$$
$$(S \oplus M)(n,k) = \max_{(\phi,\tau)\in\mathbb{R}^2} \{S(n,k) + M(n-\phi, k-\tau)\},$$

where (ϕ, τ) ranges over the domain of definition of M.

There are two possible operators generated by the combination of erosion and dilation using the same structuring element for both operations: opening $(S \circ M)$ and closing $(S \bullet M)$. The first one is an erosion followed by a dilation and the second, a dilation followed by an erosion. Mathematically it can be expressed as:

$$S \circ M = (S \ominus M) \oplus M, \tag{9}$$
$$S \bullet M = (S \oplus M) \ominus M. \tag{10}$$

The opening operator tends to remove the outer tiny leaks and round shapes, whereas the closing operator preserves the regions that have a similar shape as the structuring element. Previous experiments [32] show that closing performs better for ASR than opening.

For producing the final *masked* cochleogram S', first the closing operator is applied on the original (possibly de-noised) spectro-temporal representation S using the structuring element M and the result is subsequently added on S:

$$S' = \lambda S + (1-\lambda) S \bullet M. \tag{11}$$

where λ is a configuration parameter that weights both contributions and that has been set to 0.5 in our experiments ($\lambda = 1$ indicates no morphological filtering and corresponds with our baseline system). From this enhanced cochleogram S', mel-frequency or power-normalized based coefficients are computed following the procedure explained in Section IV-C and represented in Fig. 1.

B. Structuring Element

In this section we describe auditorily motivated structuring elements that try to emulate the complex phenomenon of cochlear masking when used in combination with MF. The SE acts as the cochlea's response to tone maskers, and the morphological filtering mechanism reproduces the masking itself. Three different structuring elements are presented, the *piecewise-linear*, *piecewise-paraboloid* and *piecewise-convex* models.

The basic piecewise-linear model for masking can be observed in Fig. 2(a) (continuous blue line). This SE is built with linear slopes for the simultaneous masking model and the logarithmic model of Equation (8) for the temporal masking. In this model, referred to as the *idealized model of masking* in Section III-C, the SE for a single frequency-time point at (n, k) is not smooth.

To be consistent with the smoothness constraint we created two new SE based in 3D quadrics, built by aggregating 4 asymmetric quadric quadrants of different parameters centered at (n, k) fitted to the empirical models in Sections III-A and III-B.

The piecewise-paraboloid model is built by aggregating paraboloid quadrants and the piecewise-convex model using hyperboloid quadrants. A comparison of the masking response of these models with the piecewise-linear model projected onto the time and frequency coordinates can be observed in Fig. 2.

As confirmed by the results in Section V, filtering with the piecewise-convex obtains the best performance. Different sizes in both frequency and time scale were tested, and the best performance was obtained by taking 10 ms of premasking, 150 ms of postmasking, and 6 bands (in Bark scale) in frequency. The 3D shape of this structuring element can be seen in Fig. 3.

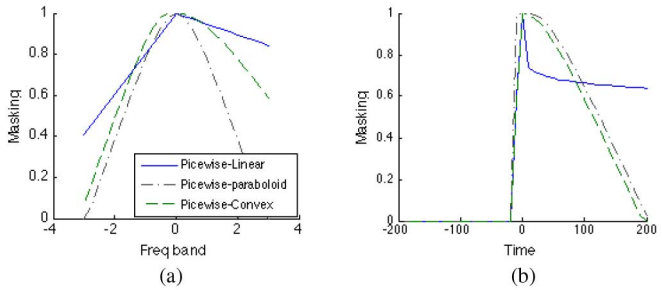

Fig. 2. Comparison between the piecewise-linear, piecewise-paraboloid and piecewise-convex models in both frequency (left) and time (right) axes. (a) Simultaneous masking (b) Temporal masking.

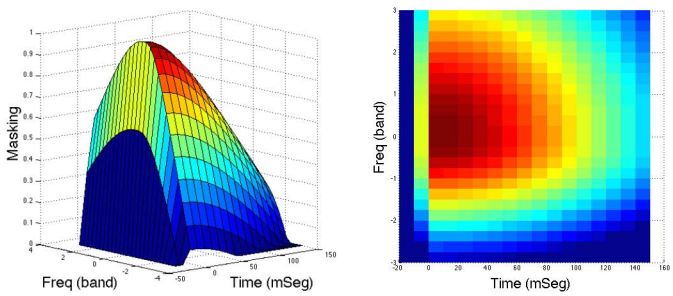

Fig. 3. Three-dimensional representation of the piece-wise convex SE. Color represents the weight of each pixel in the morphological operations. Note how temporal and simultaneous masking are interpolated by the quadrics over the parameters suggested by the pure temporal and frequency models mentioned in Section III. The asymmetry in the slopes towards higher and lower frequencies—already used in [32]—reflects the choice of different parameters to define the hyperboloids in each quadrant. This effect is more evident in the post-masking than in the pre-masking part of the SE's skirt.

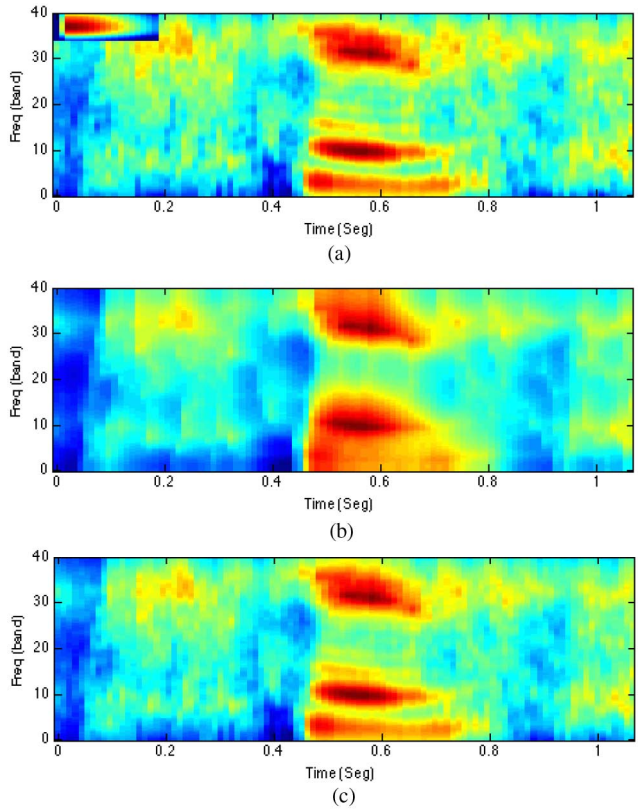

Fig. 4. Choice spectrograms output by each step of the architecture. (a) Noisy Spectrogram S compared with the SE (upper left). (b) Spectrogram after morphological filtering, $S \bullet M$. (c) Final cochleogram S' with $\lambda = 0.5$.

Since the cochlear masking model is defined in terms of the Bark scale but the spectro-temporal representations considered in this work are related to the Mel (MFCC) or ERB (PNCC) scales, the appropriate transformations between scales are applied before the morphological processing. Finally, a normalization between zero and one was applied on the intensity dimension and the SE was padded with zeros in the negative time region to center it in the mask around the pixel in which the morphological closing operation is to be performed.

The SE finally chosen can be seen at the upper left of Fig. 4(a) at scale, along with examples of the output of some of the processing steps leading to the final cochleogram.

C. Morphological Filtering-based Front-ends

In this subsection, we describe how the morphological filtering is embedded in the whole feature extraction process for automatic speech recognition.

Fig. 1 represents the block diagram of the two complete proposed front-ends based on Mel-frequency (left) and Power Normalized (right) spectro-temporal representations where the shadow blocks are our additions to, respectively, conventional MFCC and PNCC feature extraction: MF and SS. What we call a *masked cochleogram*, $S'(f,t)$, is obtained by performing morphological filtering on $S(f,t)$ using one of the single structuring elements described in Section IV-B. As for the spectral subtraction block, we found synergies with MF under the MFCC framework in our previous work [32], [33], [49] that we also confirm in this paper for PNCC (see Section V). The last two blocks in both schemes carry out the usual procedure, to de-correlate the resulting filter-bank energies by means of the Discrete Cosine Transform (DCT), followed by a Mean and Variance Normalization (MVN).

V. Experimental Results

In this section we present the experiments carried out on three different datasets: Aurora 2, ISOLET and a noisy contaminated version of Wall Street Journal.

A. Feature Extraction

As mentioned before, two different spectro-temporal representations were considered: mel-frequency and power-normalized cochleograms (see Section II). For either type, speech was analyzed using a frame length of 25 ms and a frame shift of 10 ms. After preemphasis and Hamming windowing an auditory filter bank analysis was applied over the spectrogram computed by using the Short-Time Fourier Transform (STFT). In particular, in the case of the mel-frequency representation, a set of triangular mel-scaled filters were used, whereas, for power-normalized cochleograms a bank of 40 gammatone-shaped filters whose center frequencies are linearly spaced in the ERB scale between 200 Hz and 4000 Hz was applied, followed by the PNCC [16] medium-duration power bias subtraction and power function nonlinearity. In both cases, in order to decorrelate the filterbank log-energies obtained in the previous stage, a DCT was computed over them. Cepstral coefficients C_0 to C_{12} were retained together with their corresponding delta (Δ) and acceleration ($\Delta\Delta$) coefficients to yield feature vectors of 39 components. Mean and variance normalizations were applied on each of the components.

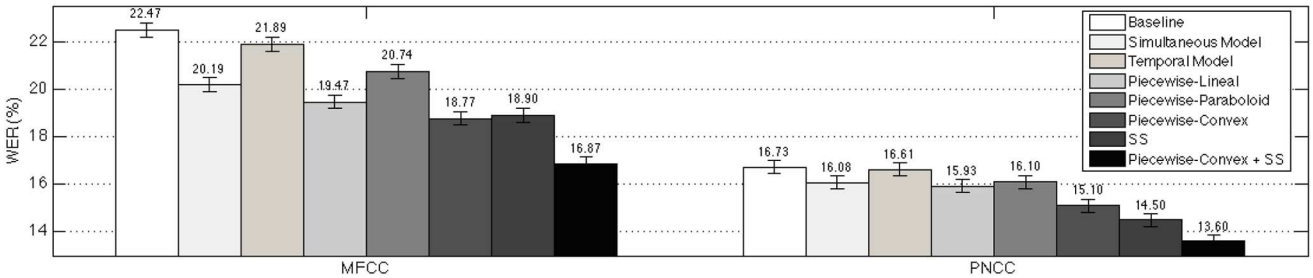

Fig. 5. Recognition results in terms of WER[%] and 95% confidence intervals using the AURORA 2 dataset (averaged over all sets).

When indicated, a conventional SS was employed over the noisy signal in order to emphasise speech over noise and MF applied over the corresponding enhanced cochleograms. Samples of the features files for the different datasets, and the scripts to replicate the results on the Aurora 2 dataset are available at [50].

B. AURORA 2 Dataset

We used the AURORA 2 dataset [51], to test our model, and to select the best structuring element. In particular, the proposed front-ends were tested in mismatched conditions.

AURORA 2 consists of a set of connected digits spoken by American English speakers and recorded at a sample rate of 8 KHz. The database was contaminated with a selection of 8 different real-world noises (subway, babble, car, exhibition hall, restaurant, street, airport and train station) at different Signal-to-Noise Ratios (SNR). In particular, SNRs from 0 dB to 20 dB with 5 dB step were considered for our experimentation. The recognizer was based on HTK (Hidden Markov Model Toolkit) software package [52] with the configuration included in the standard experimental protocol of the database described in [51], where a standard Gaussian Mixture Model (GMM)-HMM system with a 16-state word-based HMM and a 5-state silence model was adopted. As our system was tested in mismatched conditions, acoustic models were obtained from the clean training set of the database, whereas test files correspond to the complete test sets A, B and C.

Recognition results in terms of Word Error Rate (WER) and their 95% confidence intervals are shown in Fig. 5. These results correspond to several experiments carried out to study the impact of MF with the SE described in Section IV applied in isolation or in combination with SS and employing mel-frequency or power-normalized based spectro-temporal representations (labeled respectively, as *MFCC* and *PNCC*).

We consider first the influence of MF in ASR system performance with different SEs. From Fig. 5, applying MF only in the frequency domain to simulate simultaneous masking (results labeled as *Simultaneous Model*) produces better results than applying MF only in the temporal domain (results labeled as *Temporal Model*). The comparison between the three three-dimensional SE considered (*piecewise-linear*, *piecewise-paraboloid* and *piecewise-convex*) indicates that the last one outperforms the other 3D models as well as the baseline and the simultaneous and the temporal models for both spectro-temporal representations and therefore was chosen for the subsequent experiments. In particular, the application of MF with the *piecewise-convex* SE over noisy spectrograms produces relative error reductions of 16.5% for *MFCC* and 9.7% for *PNCC* with respect to the corresponding baselines, both statistically significant. This suggests that the proposed model is suitable for representing the robust behavior of the HAS in the presence of noise.

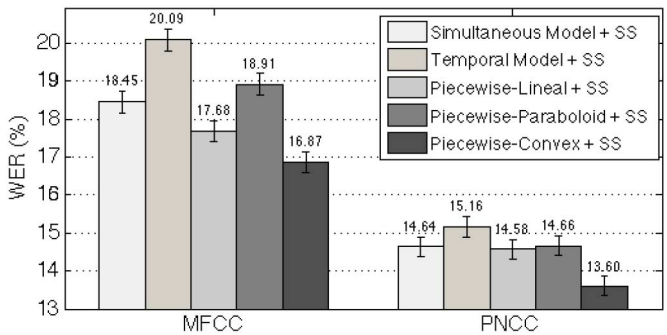

Fig. 6. Recognition results in terms of WER[%] and 95% confidence intervals using the AURORA 2 dataset (average over all the sets) for the different structuring elements in combination with spectral subtraction (SS).

Furthermore, Fig. 6 presents the results obtained employing the different proposed SE in combination with SS: the piecewise-convex SE obtained the best performance using either MFCC or PNCC. For these reasons, from now on in this paper, MF will refer to morphology filtering with the *piecewise-convex* SE.

Secondly, we also investigated combinations of SS and MF. As expected, for both spectro-temporal representations, SS with no MF clearly outperforms the corresponding baselines. For both, *MFCC* and *PNCC*, the joint use of SS and MF improves the recognition rates obtained with SS in a statistically significant manner. In particular, for *MFCC* the relative error reduction achieved by MF+SS with respect to SS is 10.7% and 24.9% with respect to the baseline. The relative error reduction obtained with *PNCC* is 6.2% and 18.7% related to using only SS and the baseline, respectively. These results show that a positive synergy exists between the SS and MF techniques. Other spectral suppression methods like MMSE [31] and Wiener [30] filtering were also initially tested but yielded worse results than SS in conjunction with PNCC.

Third, the comparison of both spectro-temporal representations shows that the different versions of features based on *PNCC* (baseline, SS, MF, SS+MF) achieve in all cases better recognition rates than the corresponding features based on *MFCC*. The best combination of *PNCC* (MF+SS) produces a relative error reduction of 19.4% with respect to the best combination of *MFCC* (MF+SS) and of 39.5% with respect to the *MFCC* baseline. Also, it is worth noting that even PNCC in isolation obtains similar results than the best combination of *MFCC*-based features (MF+SS).

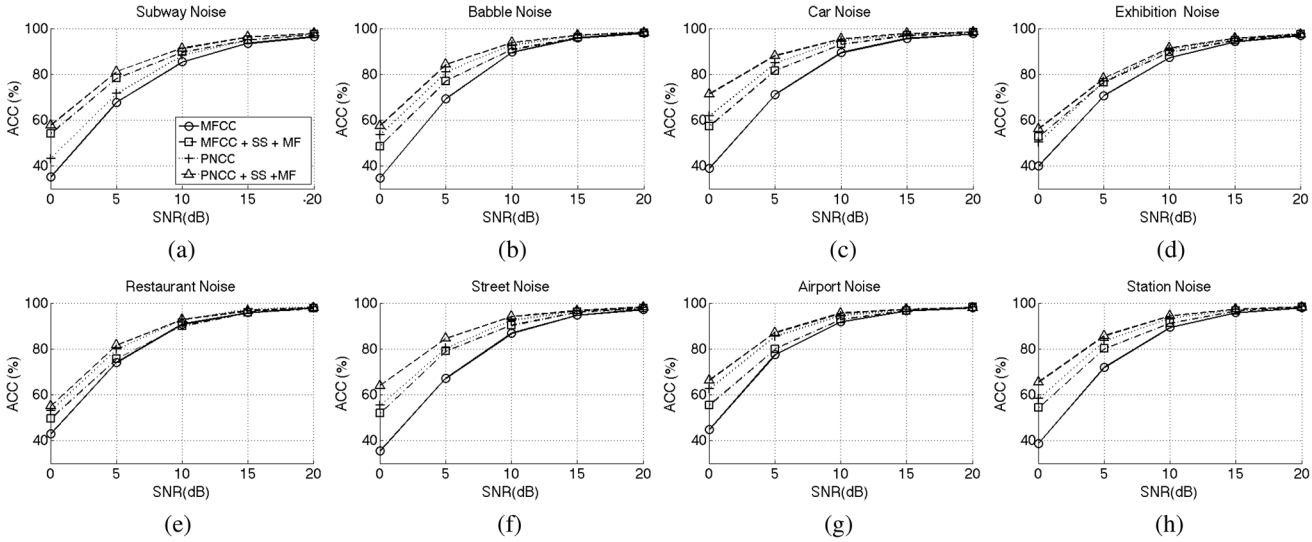

Fig. 7. Recognition results obtained under different additive noise conditions in terms of ACC[%] using the AURORA 2 dataset.

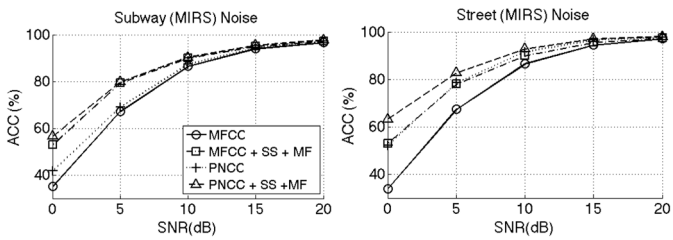

Fig. 8. Recognition results obtained under different convolutive noise conditions in terms of ACC[%] using the AURORA 2 dataset.

Fig. 7 and Fig. 8 show the recognition Accuracy (ACC) for each type of noise and SNR. For brevity's sake, only the results obtained by the baselines and MF+SS are shown in these figures. It can be observed that the *PNCC* (MF+SS) method achieves the best performance in almost every noise and SNR conditions. In some cases the *MFCC* method (MF+SS) achieves similar results to *PNCC* as can be gleaned from Figs. 7(d), (f), (h) and Fig. 8(b). Results in the presence of convolutional noise as in Fig. 8 show no degradation compared to the results obtained in the presence of additive noise only.

To conclude, we have achieved a better relative error reduction in the AURORA 2 database than some other state-of-the-art techniques; for instance, 2D-Gabor features based on power-normalized spectrograms achieve a relative error reduction of only 7.04% compared to PNCC using a HMM back-end [23].

C. ISOLET Dataset

In this section, we present the experiments carried out on the ISOLET database [53]. This database consists of 7 800 English alphabet spoken letters (two productions of each letter per each of the 150 speakers) at a sample rate of 16 KHz. Specifically, we used a version of this database called noisy-ISOLET [54] where the original ISOLET was contaminated with 8 different noise types from the NOISEX database at several SNRs (clean, 0, 5, 10, 15 and 20 dB). The noise types are: speech babble, factory noises 1 and 2, car, pink, F-16 cockpit, destroyer operations room and military vehicle noise.

The experiments were performed using the ISOLET testbed described in [54]. In particular, we trained a hybrid MultiLayer Perceptron (MLP)-HMM system [55] using a context of 5 frames to yield an MLP input dimension of 195 and only one hidden layer is employed. We employed the Quicknet multi-layer perceptron (MLP) package for acoustic modeling [56].

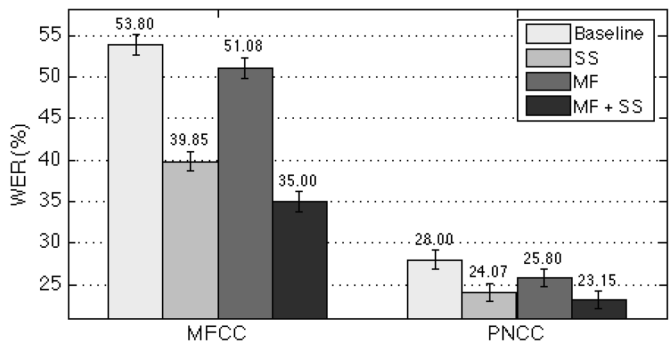

Fig. 9. Recognition results in terms of WER[%] and 95% confidence intervals using the ISOLET dataset (average over all the noises and SNR, tested in mismatched conditions).

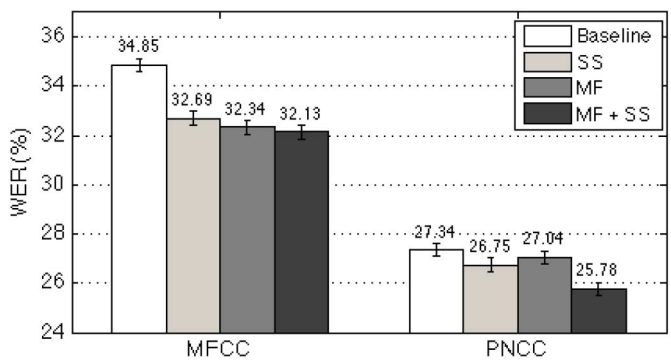

Fig. 10. Recognition results in terms of WER[%] and 95% confidence intervals using a noisy contaminated version of the WSJ0 dataset (average over all the noises and SNR).

This system was tested in *mismatched* conditions where the system is trained using clean speech and the test set consists of speech contaminated with a balanced combination of the previously mentioned noises at several SNRs. A 5-fold leave-one-out procedure was used to improve the statistical significance of the results. The corresponding recognition results in terms of WER and their 95% confidence intervals are shown in Fig. 9.

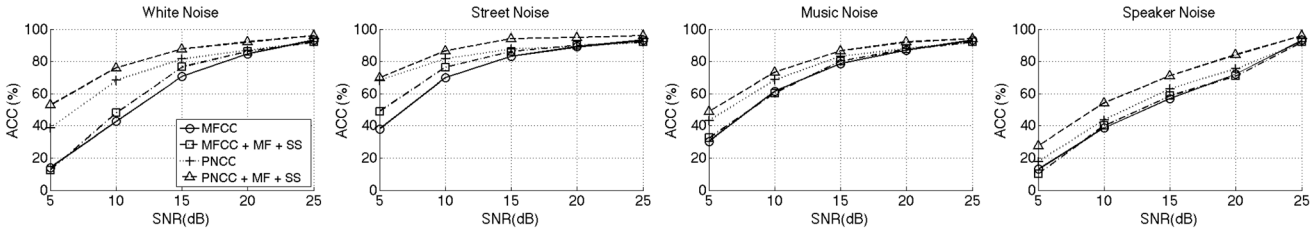

Fig. 11. Recognition results obtained under different additive noise conditions in terms of ACC[%] using the WSJ0 dataset.

We obtained similar results to those using the AURORA 2 dataset, as can be seen in Fig. 9 where, first, SS alone (without MF) clearly outperforms the corresponding baselines for both types of spectro-temporal representation (*MFCC* and *PNCC*-based). Second, the application of MF increases the recognition rates with respect to the corresponding baselines for both representations. Third, the joint use of SS and MF improves the recognition rates obtained with SS in a statistically significant manner. And last, the *PNCC* features (baseline, SS, MF, SS+MF) achieve in all cases better recognition rates than the corresponding features based on *MFCC*.

With this set of experiments we have shown that the proposed front-ends achieve also good results in hybrid ASR systems. Besides, in comparison with our previous work over the ISOLET database [32], it can be observed that we have successfully improved the design of the three-dimensional SE by means of the incorporation of perceptual facts, yielding better results.

D. WSJ0 Dataset

In this section, we present the experiments carried out on the Wall Street Journal (WSJ) database, consisting of read speech from a machine-readable corpus of WSJ news text [57]. The experiments were performed using the Hidden Markov Model Toolkit (HTK) recipe described in [58], employing a tri-gram language model with 5k vocabulary size and the Carnegie Mellon University (CMU) pronunciation dictionary.

To test the robustness of the different methods we used the same four standard testing environments as [17]: (1) white noise, (2) noise recorded live on urban streets, (3) single-speaker interference and (4) background music. The street noise was recorded on streets with steady but moderate traffic. The masking signal used for single-speaker-interference experiments consisted of other utterances drawn from the same database as the target speech, and background music was selected from music segments from the original Defense Advanced Research Projects Agency (DARPA) Hub 4 Broadcast News database.

For training the acoustic models, we used the WSJ0 SI-84 training set which contains 7 308 clean recordings (14 h). The different front-ends were tested on noisy versions of the WSJ0 5 K test set, obtained by digitally adding the previously mentioned noises—white, street, speaker and music—to the corresponding clean speech at four different SNRs using the FaNT tool [59] (with G.712 filtering). All the noise tests are evaluated in mismatched conditions (that is, training on clean speech and testing on noisy speech).

Recognition results in terms of WER and their 95% confidence intervals are shown in Fig. 10. These results correspond to the average over all the noises and SNR conditions outlined above. The performances of our systems on clean speech employing the WSJ0 5 K test set are: 5.36% WER for MFCC and 6.67% WER for PNCC.

Fig. 11 shows the recognition Accuracy (ACC) for each type of noise and SNR. For brevity's sake only the results obtained by the baselines and MF+SS are shown in these figures.

Fig. 10 shows that: (1) The *PNCC* spectral representation baseline clearly outperforms the corresponding *MFCC* baseline; (2) the application of MF improves the baseline recognition rates but not in a significant way for the *PNCC* case; (3) the joint use of SS and MF improves the recognition rates obtained with SS and with the baseline in a statistically significant manner for both representations; (4) the *PNCC* (baseline, SS, MF, SS+MF) achieve in all cases better recognition rates than the corresponding features based on *MFCC*, and (5) the improvements in the WSJ0 dataset are lower than the AURORA and ISOLET datasets. We suggest that this reduction is due to the larger size of the database and the influence of the language models in the acoustic decoding process.

Note also, from Fig. 11, that the *PNCC* (MF+SS) method achieves the best performance in every noise and SNR conditions. The improvement in white noise in Fig. 11(a), and speaker noise in Fig. 11(c) conditions are particularly worth noticing, since the proposed method clearly outperforms the PNCC baseline.

TABLE I
AVERAGE RUNTIME PER UTTERANCE FOR THE DIFFERENT METHODS OVER ALL TESTING SETS ON THE AURORA 2 DATASET

Method	Time (ms)	% from baseline
MFCC Baseline	19.66	–
MFCC + SS	26.98	37.23 %
MFCC + MF	21.84	11.80 %
MFCC + MF + SS	28.82	46.59 %
PNCC Baseline	67.93	–
PNCC+ SS	85.69	26.14 %
PNCC + MF	69.45	2.23 %
PNCC + MF + SS	87.06	28.16 %

E. Computational Complexity

Table I shows a comparison of the runtime for the different methods under different conditions (clean and noisy speech), using a workstation with 3.4 GHz Intel Core i7 and 16 GB of RAM memory. The running times were obtained by averaging each of the utterances over all testing sets on the AURORA 2 dataset. The extra time added by MF is relatively low for either MFCC or PNCC. It is worth noting that the time spent by MFCC + MF + SS is below the PNCC baseline, despite obtaining similar results in almost every noisy condition.

VI. Conclusions and Further Work

In this paper we present an enhanced, perceptually-motivated SE for morphological filtering of speech that models the complexity of HAS masking properties. Well-known empirical data in either temporal or frequency domains were interpolated to produce a three-dimensional SE for morphological filtering. A smoothness constraint was imposed since this is more suited for our hypothesis that the morphological closing operation produces a convexification of the spectro-temporal envelope of speech that models the masking properties of the HAS.

Despite ingrained intuitions that this imitation of auditory masking degrades the quality of the extracted features producing a *blurring* effect, the results we have obtained indicate that it could be in fact a sophisticated mechanism for selecting the most important parts of the spectrum from an intelligibility point of view, taking away irrelevant information and emphasizing the most robust parts of the spectrum.

The application of morphological processing with this SE in conjunction with the Power-Normalized spectro-temporal representation produces a significant increase in recognition rates in Aurora 2, ISOLET and a noisy contaminated version of the Wall Street Journal datasets. Also the results show that our method improves the recognition rates in both hybrid and traditional HMM based back-ends. To reach these results we have tested the combination of PNCC, spectral subtraction and morphological processing.

Future work will focus on the introduction of the dependency of the masker strength into morphological filtering and its interaction with alternative acoustic models such as those based on Deep Neural Networks (DNN).

Acknowledgment

We would like to thank Chanwoo Kim for kindly providing the testing noises for the WSJ0 dataset and PNCC source code.

References

[1] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.

[2] H. Fletcher and W. A. Munson, "Loudness, its definition, measurement and calculation," *J. Acous. Soc. Amer.*, vol. 5, no. 2, pp. 82–108, Oct. 1933.

[3] H. Hermansky, B. Hanson, and H. Wakita, "Perceptually based linear predictive analysis of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Apr. 1985, vol. 10, pp. 509–512.

[4] D.-S. Kim, S.-Y. Lee, and R. Kil, "Auditory processing of speech signals for robust speech recognition in real-world noisy environments," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 1, pp. 55–69, Jan. 1999.

[5] A. Ali, J. Van der Spiegel, and P. Mueller, "Robust auditory-based speech processing using the average localized synchrony detection," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 279–292, Jul. 2002.

[6] U. H. Yapanel and J. H. Hansen, "A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition," *Speech Commun.*, vol. 50, no. 2, pp. 142–152, 2008.

[7] F. Müller and A. Mertins, "Contextual invariant-integration features for improved speaker-independent speech recognition," *Speech Commun.*, vol. 53, no. 6, pp. 830–841, 2011.

[8] N. Moritz, J. Anemuller, and B. Kollmeier, "Amplitude modulation spectrogram based features for robust speech recognition in noisy and reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2011, pp. 5492–5495.

[9] A. Fazel and S. Chakrabartty, "Sparse auditory reproducing kernel (spark) features for noise-robust speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1362–1371, May 2012.

[10] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct 1994.

[11] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing," in *Readings in Speech Recognition*, A. Waibel and K.-F. Lee, Eds. San Francisco, CA, USA: Morgan Kaufmann, 1990, pp. 101–111.

[12] O. Ghitza, "Auditory nerve representation as a front-end for speech recognition in a noisy environment," *Comput. Speech Lang.*, vol. 1, no. 2, pp. 109–130, Dec. 1986.

[13] J. Allen, "Cochlear modeling," *ASSP Magazine, IEEE*, vol. 2, no. 1, pp. 3–29, Jan. 1985.

[14] R. Lyon, "A computational model of filtering, detection, and compression in the cochlea," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 1982, vol. 7, pp. 1282–1285.

[15] R. Lyon, "A computational model of binaural localization and separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Apr. 1983, vol. 8, pp. 1148–1151.

[16] C. Kim and R. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Mar. 2012, pp. 4101–4104.

[17] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2012, pp. 4101–4104.

[18] M. G. Heinz, X. Zhang, I. C. Bruce, and L. H. Carney, "Auditory nerve model for predicting performance limits of normal and impaired listeners," *Acoust. Res. Lett. Online*, vol. 2, no. 3, pp. 91–96, 2001.

[19] K. Paliwal and B. T. Lilly, "Auditory masking based acoustic front-end for robust speech recognition," in *Proc. IEEE TENCON '97. IEEE Region 10 Ann. Conf. Speech Image Technol. for Comput. Telecomm.*, 1997, vol. 1, pp. 165–168, 1.

[20] Y. Hu and P. Loizou, "Incorporating a psychoacoustical model in frequency domain speech enhancement," *IEEE Signal Process. Lett.*, vol. 11, no. 2, pp. 270–273, Feb. 2004.

[21] S. Haque, "Utilizing auditory masking in automatic speech recognition," in *Proc. Int. Conf. Audio Lang. Image Process. (ICALIP)*, 2010, pp. 1758–1764.

[22] B. T. Meyer and B. Kollmeier, "Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition," *Speech Commun.*, vol. 53, no. 5, pp. 753–767, 2011.

[23] B. T. Meyer, C. Spille, B. Kollmeier, and N. Morgan, "Hooking up spectro-temporal filters with auditory-inspired representations for robust automatic speech recognition," in *Proc. Interspeech (13th Int. Conf. Speech Commun. Technol.)*, 2012.

[24] B. T. Meyer, S. V. Ravuri, M. R. Schdler, and N. Morgan, "Comparing different flavors of spectro-temporal features for ASR," in *Proc. Interspeech (12th Int. Conf. Speech Commun. Technol.)*, 2011, pp. 1269–1272.

[25] M. Heckmann, X. Domont, F. Joublin, and C. Goerick, "A hierarchical framework for spectro-temporal feature extraction," *Speech Commun.*, vol. 53, no. 5, pp. 736–752, 2011.

[26] A. Hurmalainen and T. Virtanen, "Modelling spectro-temporal dynamics in factorisation-based noise-robust automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2012, pp. 4113–4116.

[27] C. Martínez, J. Goddard, D. Milone, and H. Rufiner, "Bioinspired sparse spectro-temporal representation of speech for robust classification," *Comput. Speech Lang.*, vol. 26, no. 5, pp. 336–348, 2012.

[28] R. Stern and N. Morgan, "Hearing is believing: Biologically inspired methods for robust automatic speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 34–43, Nov. 2012.

[29] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1979, vol. 4, pp. 208–211.

[30] P. Scalart and J. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1996, vol. 2, pp. 629–632.

[31] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, 1984.

[32] J. Cadore, F. J. Valverde-Albacete, A. Gallardo-Antoln, and C. Peláez-Moreno, "Auditory-inspired morphological processing of speech spectrograms: Applications in automatic speech recognition and speech enhancement," *Cognitive Comput.*, vol. 5, no. 4, pp. 426–441, 2013.

[33] J. Cadore, A. Gallardo-Antolín, and C. Peláez-Moreno, "Morphological processing of spectrograms for speech enhancement," in *Advances in Nonlinear Speech Processing*. Berlin, Germany: Springer-Verlag, 2011, pp. 224–231.
[34] J. Hansen, "Morphological constrained feature enhancement with adaptive cepstral compensation (MCE-ACC) for speech recognition in noise and Lombard effect," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 598–614, Oct. 1994.
[35] H. Fastl and E. Zwicker, *Psycho-acoustics: Facts and Models*, 3rd ed. New Yotk, NY, USA: Springer, 2007.
[36] F. de-la Calle-Silos, F. J. Valverde-Albacete, A. Gallardo-Antolín, and C. Peláez-Moreno, "ASR feature extraction with morphologically-filtered power-normalized cochleograms," in *Proc. Interspeech (15th Int. Conf. Speech Commun. Technol.)*, 2014, pp. 2430–2434.
[37] E. Zwicker and E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," *J. Acoust. Soc. Amer.*, vol. 68, p. 1523, 1980.
[38] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude of pitch," *J. Acoust. Soc. Amer.*, vol. 8, pp. 185–190, 1937.
[39] B. Glasberg and B. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 47, no. 1–2, pp. 103–138, 1990.
[40] B. Moore and B. Glasberg, "A revised model of loudness perception applied to cochlear hearing loss," *Hear. Res.*, vol. 188, no. 1–2, pp. 70–88, 2004.
[41] B. Moore and B. Glasberg, "Suggested formula for calculating auditory-filter bandwidths and excitation patterns," *J. Acoust. Soc. Amer.*, vol. 74, p. 750, 1983.
[42] R. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," *Auditory Physiol. Percept.*, vol. 83, pp. 429–446, 1992.
[43] R. D. Patterson, K. Robinson, J. Holdsworth, D. Mckeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Proc. 9th Int. Symp. Hear. Audit., Physiol. Percept.*, 1992, pp. 429–446.
[44] G. V. Békésy, "On the resonance curve and the decay period at various points on the cochlear partition," *J. Acous. Soc. Amer.*, vol. 21, no. 3, pp. 245–254, 1949.
[45] E. Zwicker and A. Jaroszewski, "Inverse frequency dependence of simultaneous tone-on-tone masking patterns at low levels," *J. Acous. Soc. Amer.*, vol. 71, no. 6, pp. 1508–1512, 1982.
[46] W. Jesteadt, S. P. Bacon, and J. Lehman, "Forward masking as a function of frequency, masker level, and signal delay," *J. Acous. Soc. Amer.*, vol. 71, no. 4, pp. 950–962, 1982.
[47] G. Matheron and J. Serra, "The birth of mathematical morphology," in *Proc. 6th Int. Symp. Math. Morphol.*, Sydney, Australia, 2002, pp. 1–16.
[48] E. R. Dougherty and R. A. Lotufo, *Hands-On Morphological Image Processing*, ser. Tutorial Texts in Optical Engineering. Bellingham, WA, USA: SPIE Press, 2003.
[49] J. Cadore, C. Peláez-Moreno, and A. Gallardo-Antolín, "Morphological processing of a dynamic compressive gammachirp filterbank for automatic speech recognition," *IberSPEECH'12*, 2012.
[50] F. de la Calle Silos, "Source code," [Online]. Available: http://www.tsc.uc3m.es/fsilos
[51] D. Pearce and H.-G. Hirsch, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ICSLP 6th Int. Conf. Spoken Lang. Process.*, Oct. 2000, no. , pp. 16–19.
[52] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK book, ver 3.4.*. Cambridge, U.K.: Entropic Cambridge Res. Lab., 2006.
[53] K. Bache and M. Lichman, *The Isolet Spoken letter database* [Online]. Available: https://archive.ics.uci.edu/ml/datasets/ISOLET, last accessed: 2015-07-02
[54] D. Gelbart, H. W., M. Holmberg, and N. Morgan, *Noisy isolet and isolet testbeds,* [Online]. Available: http://www1.icsi.berkeley.edu/Speech/papers/eurospeech05-onset/isolet/, last accessed: 2015-07-02
[55] H. Bourlard and N. Morgan, "Hybrid HMM/ANN systems for speech recognition: Overview and new research directions," *Adaptive Process. Sequences and Data Structures*, pp. 389–417, 1998.
[56] D. Johnson, ICSI quicknet software package. [Online]. Available: http://www.icsi.berkeley.edu/Speech/qn.html last accessed: 2015-07-02,
[57] D. B. Paul and J. M. Baker, "The design for the wall street journal-based csr corpus," in *Proc. Workshop Speech Nat. Lang.*, Stroudsburg, PA, USA, 1992, pp. 357–362 [Online]. Available: http://dx.doi.org/10.3115/1075527.1075614, ser. HLT '91., Assoc. for Comput. Linguist.
[58] K. Vertanen, "Baseline WSJ acoustic models for HTK and Sphinx: Training recipes and recognition experiments," Cavendish Lab., Cambridge, U.K., Tech. Rep., 2006.
[59] G. Hirsch, *FaNT - Filtering and Noise Adding Tool*, 2005 [Online]. Available: http://dnt.kr.hsnr.de/download.html

Fernando de la Calle Silos received his B.Sc. in audiovisual system engineering in 2012 and his M.Sc. in multimedia and communications in 2014 both from Universidad Carlos III de Madrid (Spain). He is currently working toward a Ph.D. in the same university. His main research interests include automatic speech recognition and signal processing for multimedia human–machine interaction. He has been working in computer vision a few years after beginning his Ph.D.

Francisco J. Valverde-Albacete received his Eng. degree in telecommunications from Universidad Politécnica de Madrid (Spain) in 1992 and his D.Eng. in telecommunications from Universidad Carlos III de Madrid (Spain) in 2002.

Before joining UNED in 2013, he was Associate Professor at the Signal Theory and Communications Department, Universidad Carlos III de Madrid. He has also visited with U. of Strathclyde, U. degli Study di Trento and the International Computer Science Institute (ICSI, Berkeley).

Dr. Valverde-Albacete has published over 50 papers in books, journals and conferences in applied maths, speech and language processing, machine learning and data mining, where his interests lie. He is a member of IEEE and ACM.

Ascensión Gallardo-Antolín received her Ph.D. in telecommunication engineering from the Polytechnic University of Madrid, Spain, in 2002. She has been a Visiting Scientist at the International Computer Science Institute (ICSI, Berkeley, USA) in 2005, the German Research Center for Artificial Intelligence (DFKI, Saarbrucken, Germany) in 2006 and the Centre for Speech Technology Research (CSTR, University of Edinburgh, UK) in 2013. Currently, she is an Associate Professor at the Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Spain. She has coauthored more than 70 peer-reviewed papers in international journals and national and international conferences. She has participated in several research projects including some of the Spanish Council on Science and Technology and the UE. She has received the Best Ph.D. Thesis Award from the Professional Association of Telecommunication Engineers of Spain (COIT) and the Ph.D. Excellence Award from the Polytechnic University of Madrid. Her main research interests include automatic speech recognition, audio classification and segmentation, multimedia information retrieval, auditory and visual salience models and signal processing for multimedia human–machine interaction.

Carmen Peľaez-Moreno received her Telecommunication Eng. degree from the Public University of Navarre in 1997 and Ph.D. from the University Carlos III of Madrid in 2002. She is currently an Associate Professor in the Department of Signal Theory and Communications at the University Carlos III of Madrid. She has been a visiting researcher at Strathclyde University (Glasgow, UK), the International Computer Science Institute's (Berkeley, USA) and the University of Trento (Italy).

Her research interests include speech recognition and perception, multimedia processing, machine learning and data analysis. She has co-authored over 60 papers in journals, books and peer-reviewed conferences.