ASR Feature Extraction with Morphologically-Filtered Power-Normalized Cochleograms

F. de-la-Calle-Silos[1], *F.J. Valverde-Albacete*[2], *A. Gallardo-Antolín*[1], *C. Peláez-Moreno*[1]

[1]Department of Signal Theory and Communications.
Universidad Carlos III, Leganés (Madrid), Spain
[2]Departamento de Lenguages y Sistemas Informáticos.
Univ. Nacional de Educación a Distancia, Madrid, Spain
fsilos@tsc.uc3m.es, fva@lsi.uned.es, gallardo@tsc.uc3m.es, carmen@tsc.uc3m.es

Abstract

In this paper we present advances in the modeling of the masking behavior of the Human Auditory System to enhance the robustness of the feature extraction stage in Automatic Speech Recognition. The solution adopted is based on a non-linear filtering of a spectro-temporal representation applied simultaneously on both the frequency and time domains, by processing it using mathematical morphology operations as if it were an image. A particularly important component of this architecture is the so called *structuring element*: biologically-based considerations are addressed in the present contribution to design an element that closely resembles the masking phenomena taking place in the cochlea. The second feature of this contribution is the choice of underlying spectro-temporal representation. The best results were achieved by the representation introduced as part of the Power Normalized Cepstral Coefficients together with a spectral subtraction step. On the Aurora 2 noisy continuous digits task, we report relative error reductions of 18.7% compared to PNCC and 39.5% compared to MFCC.

Index Terms: Spectro-temporal processing, Morphological filtering, Automatic speech recognition, Auditory-based features, PNCC.

1. Introduction

Machine performance in Automatic Speech Recognition (ASR) tasks is still far away from that of humans, and noisy conditions only compound the problem. Like other researchers, we hypothesize 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 some properties of the HAS to a certain extent. For example, the Mel-Frequency Cepstral Coefficients (MFCC) [1] or the Gammatone-based Coefficients (GTC) [2], result from non-linear transformations of the frequency domain, and they model a filterbank that mimics the existence of critical bands in the cochlea. Some other aspects, like the non-linear perception of sound intensity, are also present as part of these procedures. In this paper we concentrate on the modeling of the masking phenomena in the cochlea.

Other methods based on a better modeling of HAS properties—in particular the masking effect—can be found in the literature: in [3] a masking threshold as a function of frequency is computed, [4] employes several psycho-acoustical properties of human perception to define a perceptual speech excitation function and performs Spectral Subtraction (SS) [5] in the masked region, while [6] performs an estimation of the clean signal taking into account the simultaneous masking effect. On the other hand, detailed physiological models—like the one proposed in [7] based on the auditory-nerve activity–have been used in ASR [8] effectively, but the high computational cost motivated the development of simplified models that capture the essentials of auditory processing, as the one we propose next.

In this work we refine our hypothesis that morphological filtering produces a smoothing of the spectro-temporal envelope that better models the masking behavior of the cochlea. Our model filters a spectro-temporal representation of speech—sometimes referred to as *cochleogram*—as if it were an image, allowing for the simultaneous processing of both dimensions, time and frequency. The filtering procedure we propose is based on *mathematical morphology* operations, and it aims to reproduce the masking properties of the HAS. For that purpose, the mask—or in mathematical morphology terminology the *structuring element* (SE)—employed reproduces the spectro-temporal masking behavior from well-known empirical measurements in the spectral and temporal domains independently. Despite ingrained intuitions that masking deteriorates signal quality, we propound that it smoothes away some noise and artifacts. In [9, 10] we presented 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 [11], such efforts did not take into account the HAS properties.

It seems that the design of the SE is the crux of our approach. 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 [12], relying on the underlying spectro-temporal representation to accommodate these effects. The proper choice of this representation is the second leg of our feature extraction method. We have selected the one recently proposed in [13, 14] as part of the Power-Normalized Cepstral Coefficients (PNCC) in combination with conventional spectral subtraction.

PNCC includes the use of a power-law non-linearity that replaces the traditional logarithmic non-linearity used in MFCC coefficients, a noise-suppression algorithm based on asymmetric filtering that suppresses background excitation, and a mod-

ule that carries out temporal masking by placing a peak for each frequency channel and suppressing the instantaneous power if it falls below the envelope. In our work we borrow the underlying spectro-temporal representation proposed by PNCC but introduce our own temporal masking procedure that goes a step beyond that in PNCC, while still maintaining a low computational complexity.

Other solutions that simultaneously perform temporal and spectral analysis to yield spectro-temporal features have lately emerged, e.g. spectro-temporal Gabor features [15, 16, 17], HIST [18], spectro-temporal derivative features [19] or sparse spectro-temporal features [20]. Auditory-inspired representations in these domains are reviewed in [21]. Unlike [15, 16, 17], where a reduced set of temporal, spectral and spectro-temporal filters need to be chosen, in our approach a single SE is used across the board.

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 spectral subtraction, Wiener filtering [22] or the minimum mean-square error short-time spectral amplitude estimator [23]—and attempt to enhance the speech signal without extensively modeling the HAS properties.

The remainder of this paper is organized as follows: Section 2 introduces the basic terminology of mathematical morphology needed for the rest of the paper; Section 3 describes the theoretical basis for the psycho-acoustical modeling of the SE. Section 4 briefly describes the spectro-temporal representation underlying our procedure. Finally, our results are presented in Section 5 followed by some conclusions and further lines of research in Section 6.

2. An overview of morphological processing

Mathematical Morphology is a theory for the analysis of spatial structures [24] whose main application domain is in Image Processing as a tool for thinning, pruning, structure enhancement, object marking, segmentation and noise filtering [25]. It may be used on both binary and grey-scale images.

To perform Morphological Filtering (MF) operations, we first convolve the image with a SE 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 4.

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 enlargement and fill in small holes. Let S be the underlying spectro-temporal representation and M the structuring element, erosion is defined as: $S \ominus M$ and dilation: $S \oplus M$.

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:

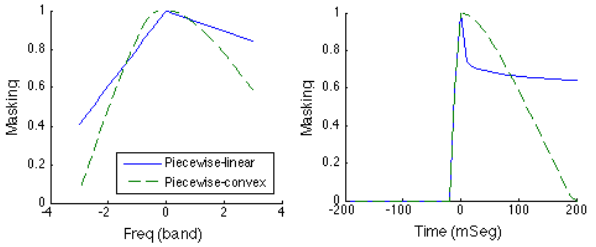
$$S \circ M = (S \ominus M) \oplus M; \quad S \bullet M = (S \oplus M) \ominus M \quad (1)$$

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 [9] show that closing performs better for ASR than opening.

For producing the final filtered 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.

$$S' = S + S \bullet M \quad (2)$$

From this enhanced cochleogram S', the cepstral coefficients are computed following the procedure explained in Section 4.



(a) Simultaneous masking (b) Temporal masking

Figure 1: Comparison between the piecewise-linear model and the proposed piecewise-convex model in both frequency (left) and time (right) scales.

3. Modeling Cochlear Masking

In this section we present a novel auditorily-motivated SE that tries to emulate the complex phenomenon of cochlear masking.

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 [26].

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. The effect of a masker on simultaneously masked frequencies is called *simultaneous masking*. The phenomenon whereby a masker affects non-simultaneous frequencies is called *temporal masking*.

Classical masking experiments concentrated in determining the amount of masking in either of these directions—frequency or time—in isolation. Such experiments, for instance, 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 [27]. 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 greater frequency maskers.

The basic *piecewise-linear* model for simultaneous masking can be observed in Figure 1.(a) (continuous blue line). It

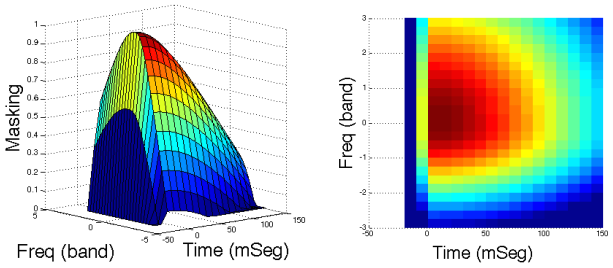


Figure 2: Visualizations of the structuring element.

consists of a piecewise linear masker threshold, with slopes of $+30$ dB per band for lower band masking and -8 dB per band for higher band masking.

Using masking tones $s(F,t) = L_m\delta(F - F_m, t - T_m)$ where F is in a transformed frequency scale, L_m is the sound pressure level of the tone, F_m and T_m are the masker frequency and time instant, the slopes were fitted for $L_m = 60$ dB (see Figure 6.14 of [12]). 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.

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 20ms prior to the masker, while the duration of postmasking extends well beyond 200 ms, perhaps as far as 500 ms [9].

Premasking is usually modeled as a constant slope of $+25$ dB/ms, starting 20 ms before the masker, and postmasking with a fitted model for single masker-induced postmasking which was presented in [28],

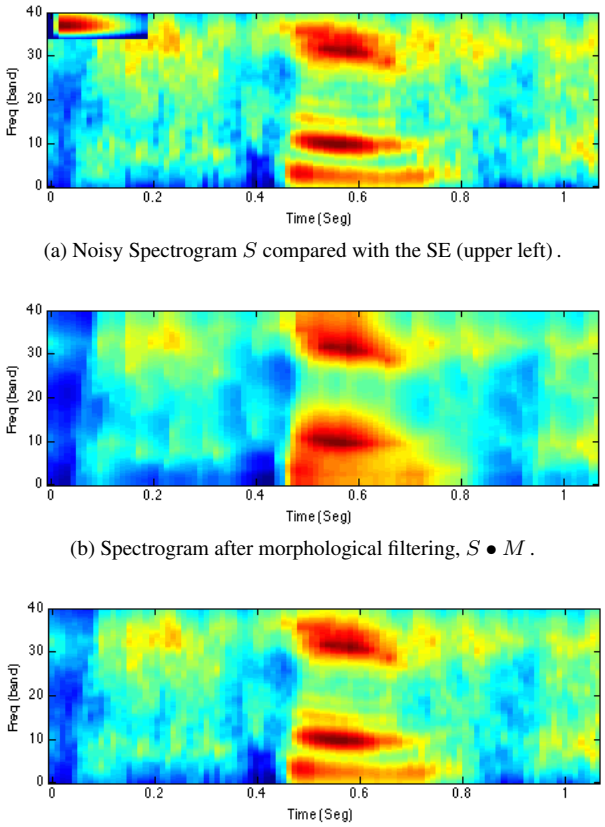
$$M\left(t - T_m, L_m\right) = a\big(b - \log\left(t - T_m\right)\big)\big(L_m - c\big) \qquad (3)$$

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. Previous premasking and postmasking models can be observed in Figure 1.(b) (continuous blue line).

After such models a masking SE for a single frequency-time point should be quite sharp. But findings point to a smooth model around (F_m, T_m), with sublinear decays close to this point and superlinear decays further away [12]. To explain this, we hypothesize that at the level of granularity at which the cochleogram is being observed the masking response of a particular (F_m, T_m) must be the aggregation of many single-point responses. Thus, we propose a piecewise-convex model built by aggregating 4 paraboloid quadrants of different parameters fitted to the contour provided by the explained time and simultaneous masking models.

The shape of the proposed SE can be seen in Figure 2. Notice the difference in the frequency slopes: this is consistent with previous work concentrating in higher frequencies and was used in [9]. A comparison with the piecewise-linear model can be observed in Figure 1 (dashed green line).

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 frequency. In all cases, the frequency resolution of each band was 4 pixels. Temporal and simultaneous masking were interpolated over these boundaries and a normalization between zero and one



(a) Noisy Spectrogram S compared with the SE (upper left) .

(b) Spectrogram after morphological filtering, $S \bullet M$.

(c) Final cochleogram S' .

Figure 3: Choice spectrograms output by each step of the architecture.

was applied. Finally, the SE was padded with zeros 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 Figure 3(a) along with examples of the output of some of the processing steps leading to the final cochleogram.

4. Spectro-temporal representation

In this section we explain our choice of the auditorily-motivated frequency scaled spectrograms or cochleograms used in the proposed front-end.

The underlying spectro-temporal representation is the domain where the previously explained MF filtering is applied. In this paper, we have chosen the power-normalized spectro-temporal representation used in the PNCC feature extraction process: ERB scale [29, 30] and a gammatone-shaped filters bank analysis, given that the impulse response of the gammatone function provides an excellent fit to the human auditory filter shapes [31] allowing a better modeling of the masking. For comparison purposes, results employing cochleograms based on the conventional mel [32] frequency scale and triangular filter bank are also presented.

The power-normalized cochleogram computation is performed as follows: the speech signal is analysed using a frame length of 25ms and a frame shift of 10ms. After preemphasis and Hamming windowing an auditory filter bank analysis is applied over the spectrogram computed by using the Short-Time Fourier Transform (STFT). In particular, a bank of 40 gammatone-shaped filters whose center frequencies are lin-

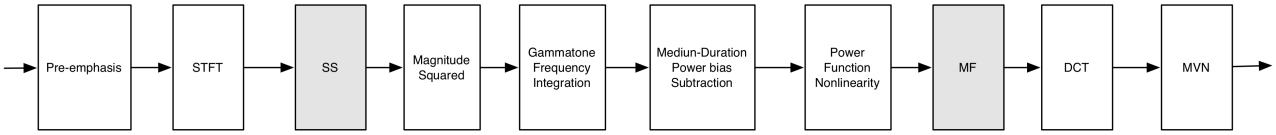


Figure 4: Structure of the porposed front-end algorithm; shaded blocks, spectral subtraction (SS) and Morphological Filtering (MF), indicates the major differences regarding the PNCC algorithm.

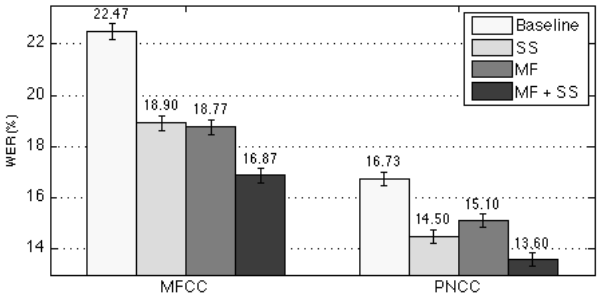


Figure 5: Recognition results in terms of WER[%] and 95% confidence intervals.

early spaced in the ERB scale between 200Hz and 4000Hz is applied, followed by the PNCC [13] medium-duration power bias subtraction and power function nonlinearity, to obtain the cochleogram S.

The *masked* cochleogram S' is obtained by performing a morphological filtering using the SE described in Section 3. Then to decorrelate the resulting filter-bank energies a Discrete Cosine Transform (DCT) is applied, followed by a mean variance normalization, to yield a modified version of the PNCC coefficients. In some of our experimental validations we add a spectral subtraction step after the STFT to show that the proposed algorithm can be combined with other noise suppression schemes. A diagram of whole process is shown in Figure 4.

5. Experimental Results

We used the AURORA 2 dataset [33], to test our model. It consists 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 at different SNRs. The experiments were performed using the HTK reference back-end described in [33], where a standard GMM-HMM system with a 16-state word-based HMM and a 5-state silence model was adopted.

Cepstral coefficients C_0 to C_{12} obtained by the proposed front-end were retained together with their corresponding delta (Δ) and acceleration ($\Delta\Delta$) coefficients to yield feature vectors of 39 components. Mean and variance normalization were applied on each of the components. The system was tested in mismatched conditions (sets A, B and C).

Recognition results in terms of Word Error Rate (WER) and their 95% confidence intervals are shown in Figure 5. These results correspond to several experiments carried out to study the impact of MF with the SE described in Section 3 applied in isolation or in combination with SS and employing MFCC- or PNCC-based spectral representations (labeled respectively, as *MFCC* and *PNCC*).

First, the influence of MF in the ASR system performance is analyzed. As can be observed, the application of MF over the 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 result suggests that the proposed model is suitable for representing the behavior of the HAS.

Second, the combination of SS and MF was also investigated. As expected, for both spectro-temporal representations, SS alone (without 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 SS and the baseline, respectively. These results show that a positive synergy exists between the SS and MF techniques.

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.

To conclude, a better relative error reduction in the AURORA 2 database was achieved than other state-of-the-art techniques. In comparison, for instance, 2D-Gabor features based on power-normalized spectrograms achieve a relative error reduction of only 7.04% with respect to PNCC using a HMM back-end [16].

6. Conclusions and further work

An enhanced biologically-motivated SE that takes into account the HAS masking properties is presented in this paper. Well-known empirical results in both temporal and frequency domains were interpolated to produce a three-dimensional SE. A smoothness restriction was imposed since this is more suited for our hypothesis that the morphological filtering produces a convexification of the spectro-temporal envelope of speech that resembles the masking properties of the HAS. The application of morphological processing in conjunction with the PNCC spectro-temporal representation produces a significant increase in recognition rates in the Aurora 2 dataset. Besides the combination of PNCC, spectral subtraction and morphological processing is investigated. Future work will focus on the introduction of the dependency of the masker strength into the morphological procedure and the application to alternative ASR architectures like hybrid and tandem approaches, that has already produced improvements with other spectro-temporal feature extraction methods.

7. Acknowledgements

This contribution has been supported by an Airbus Defense and Space Grant (Open Innovation - SAVIER) and Spanish Government-CICYT project 2011-26807/TEC.

8. References

[1] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," vol. 28, no. 4, pp. 357–366, 1980.

[2] H. Yin, V. Hohmann, and C. Nadeu, "Acoustic features for speech recognition based on gammatone filterbank and instantaneous frequency," *Speech Communication*, vol. 53, no. 5, pp. 707 – 715, 2011.

[3] K. Paliwal and B. T. Lilly, "Auditory masking based acoustic front-end for robust speech recognition," in *TENCON '97. IEEE Region 10 Annual Conference. Speech and Image Technologies for Computing and Telecommunications., Proceedings of IEEE*, vol. 1, 1997, pp. 165–168 vol.1.

[4] S. Haque, "Utilizing auditory masking in automatic speech recognition," in *Audio Language and Image Processing (ICALIP), 2010 International Conference on*, 2010, pp. 1758–1764.

[5] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Acoustics, Speech, and Signal Processing (ICASSP '79), IEEE International Conference on*, vol. 4, 1979, pp. 208–211.

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

[7] X. Zhang, M. G. Heinz, I. C. Bruce, and L. H. Carney, "A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression," *The Journal of the Acoustical Society of America*, vol. 109, no. 2, p. 648, 2001.

[8] C. Kim, Y. Chiu, and R. Stern, "Physiologically-motivated synchrony-based processing for robust automatic speech recognition." in *Interspeech*, 2006, pp. 1483–1486.

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

[10] J. Cadore, A. Gallardo-Antolín, and C. Peláez-Moreno, "Morphological processing of spectrograms for speech enhancement," *Lecture Notes in Computer Science*, pp. 224–231, 2011.

[11] J. Hansen, "Morphological constrained feature enhancement with adaptive cepstral compensation (MCE-ACC) for speech recognition in noise and Lombard effect," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 4, pp. 598–614, 1994.

[12] H. Fastl and E. Zwicker, *Psycho-acoustics: Facts and Models*, 3rd ed. Springer, 2007.

[13] C. Kim and R. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, March 2012, pp. 4101–4104.

[14] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*.

[15] 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.

[16] 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 *INTERSPEECH*, 2012.

[17] B. T. Meyer, S. V. Ravuri, M. R. Schdler, and N. Morgan, "Comparing different flavors of spectro-temporal features for ASR," in *INTERSPEECH*. ISCA, 2011, pp. 1269–1272.

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

[19] A. Hurmalainen and T. Virtanen, "Modelling spectro-temporal dynamics in factorisation-based noise-robust automatic speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 4113–4116.

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

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

[22] P. Scalart and J. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Acoustics, Speech, and Signal Processing (ICASSP'96), IEEE International Conference on*, vol. 2, 1996, pp. 629–632.

[23] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, no. 6, pp. 1109–1121, 1984.

[24] G. Matheron and J. Serra, "The birth of mathematical morphology," in *Proc. 6th Int. Symp. Mathematical Morphology*. Sydney, Australia, 2002, pp. 1–16.

[25] E. R. Dougherty and R. A. Lotufo, *Hands-on Morphological Image Processing*, ser. Tutorial Texts in Optical Engineering. SPIE press, 2003.

[26] G. v. Békésy, "On the resonance curve and the decay period at various points on the cochlear partition," *The Journal of the Acoustical Society of America*, vol. 21, no. 3, pp. 245–254, 1949.

[27] E. Zwicker and A. Jaroszewski, "Inverse frequency dependence of simultaneous tone-on-tone masking patterns at low levels," *The Journal of the Acoustical Society of America*, vol. 71, no. 6, pp. 1508–1512, 1982.

[28] W. Jesteadt, S. P. Bacon, and L. JR, "Forward masking as a function of frequency, masker level, and signal delay," *The Journal of the Acoustical Society of America*, vol. 71, no. 4, pp. 950 – 962, 1982.

[29] B. Glasberg and B. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Res.*, vol. 47, no. 1-2, pp. 103–138, 1990.

[30] B. Moore and B. Glasberg, "A revised model of loudness perception applied to cochlear hearing loss," *Hearing Research*, vol. 188, no. 1-2, pp. 70–88, 2004.

[31] R. D. Patterson, K. Robinson, J. Holdsworth, D. Mckeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *in Proc. 9th Int. Symp. Hearing Audit., Physiol. Perception*, 1992, pp. 429–446.

[32] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude of pitch," *J. Acoust. Soc. Am.*, vol. 8, pp. 185–190, 1937.

[33] D. Pearce and H.-G. Hirsch, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ICSLP 2000 ,6 th International Conference on Spoken Language Processing*, no. October, 2000, pp. 16–19.