Preliminary experiments on the robustness of biologically motivated features for DNN-based ASR

F. de-la-Calle-Silos [*], Francisco J. Valverde-Albacete [†], A. Gallardo-Antolín [*], C. Peláez-Moreno[*]
[*] Signal Theory and Communications Department,
Universidad Carlos III,
Leganés (Madrid), Spain
{fsilos, gallardo, carmen}@tsc.uc3m.es
[†]Departamento de Lenguages y Sistemas Informáticos.
Univ. Nacional de Educación a Distancia,
Madrid, Spain
fva@lsi.uned.es

***Abstract*—A perceptually motivated feature extraction method based on mimicking the masking properties of the cochlea has been recently found to provide enhanced performance when applied to conventional speech recognition back-ends. On the other hand, the introduction of Deep Neural Network (DNN) based acoustic models has produced dramatic improvements in performance. In particular, we found that Deep Maxout Networks, a modification of DNNs' feed-forward architecture that uses a max-out activation function, provides enhanced robustness to environmental noise. In this paper, we present preliminary experiments on the combination of these two elements that already show how the DMN-based back-end is capable of taking advantage of these auditorily inspired features making the whole system more robust and also suggesting that human-like representations of speech keep playing an important role in DNN-based automatic speech recognition systems.**

I. Introduction

Machine performance in Automatic Speech Recognition (ASR) tasks is still far away from that of humans, and noisy conditions only compound the problem. Noise robustness techniques can be divided into two approaches: feature enhancement and model adaptation. *Feature enhancement* tries to remove noise from the speech signal without changing the acoustic model parameters while *model adaptation* changes these parameters to fit the model to the noisy speech signal.

Our biologically inspired features [1] 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 granting a biologically meaningful application of the two previously mentioned elements.

In particular, 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, based on *mathematical morphology* operations, 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 that were either carried out in the spectral or the temporal domains separately. Thus, the design of this element is the crux of our approach. In [1], we present various structuring element designs that aim at closely resembling the auditory masking phenomena taking place in the cochlea.

Apart from these techniques, the last years have witnessed an important leap in performance with the introduction of new acoustic models based on Deep Neural Networks (DNNs) ([2], [3]).

Deep Neural Networks (DNNs) can be applied both in the so-called *tandem* [4] and *hybrid* [5] architectures. In the first case, DNNs can be trained to generate bottleneck features which are fed to a conventional GMM-HMM back-end. In the second, DNNs are employed for acoustic modeling by replacing the GMMs into an HMM system. In this paper we adopt a DNNs hybrid configuration.

However, the most remarkable difference between traditional neural networks and DNN is that a *pre-training* stage is needed to reduce the chance that the error back-propagation algorithm employed for training falls into a poor local minimum. Besides, some recent methods have been proposed to avoid overfitting and improve the accuracy of the networks, as for example, dropout [6] which randomly omits hidden units in the training stage. Another related technique is the so-called Deep Maxout Networks (DMNs) [7] that split the hidden units at each layer into non-overlapping groups, each of them generating an activation using a max pooling operation. This way, DMNs reduce the size of the parameter space significantly making it very suited for ASR tasks where the training sets and input and output dimensions are normally quite large.

In [8] we proved that DMNs improve recognition rates in noisy conditions given that they were capable to more effectively model speech variability from limited data. In this paper we combine our biologically inspired features with

978-1-4673-7846-8/15/$31.00 ©2015 IEEE

the DMNs under various noisy conditions, aimed to improve behavior in these scenarios. Achieved results show that feature engineering also has an important role in the DNN-based automatic speech recognition systems.

The remainder of this paper is organized as follows: Section II reviews the biologically inspired features, Section III introduces Deep Neural Networks, the hybrid automatic speech recognition architecture, and dropout and maxout methods. Finally Section IV presents the results obtained in the TIMIT dataset followed by some conclusions and further lines of research in Section V.

II. Morphologically-filtered biologically inspired features

A. An overview of morphological processing

Mathematical Morphology is a theory for the analysis of spatial structures [9] whose main application domain is in Image Processing as a tool for thinning, pruning, structure enhancement, object marking, segmentation and noise filtering [10]. 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 II-C.

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:

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

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 [11] 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.

$$S' = S + S \bullet M \tag{2}$$

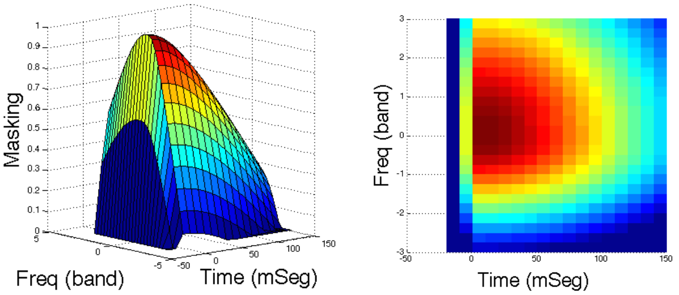


Fig. 1: Visualizations of the structuring element.

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

B. Modeling Cochlear Masking

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

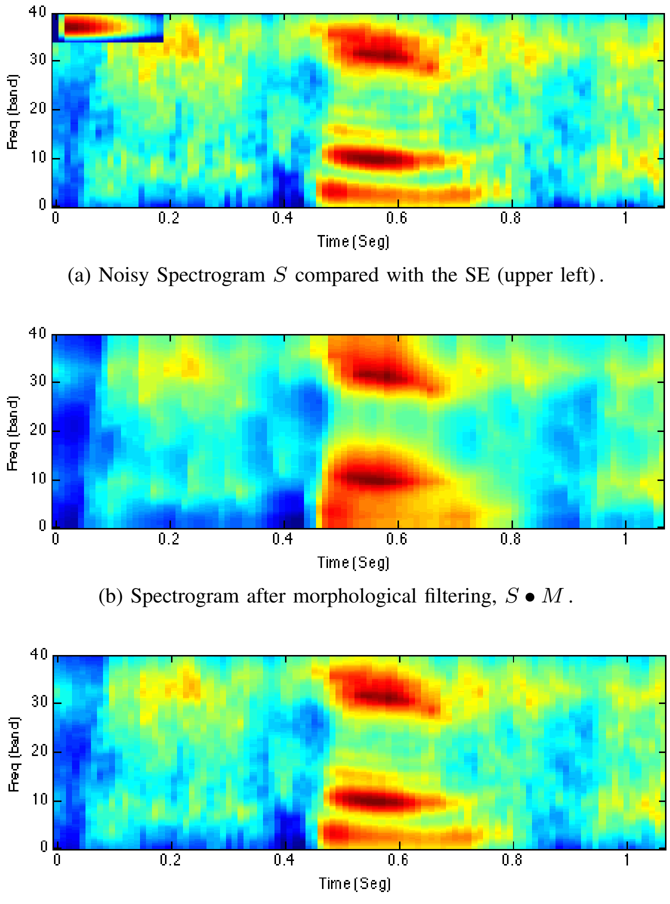
The cochlea is the organ that converts the mechanical vibrations in the middle ear to neural impulses. The basilar membrane—the sensing structure that runs the length of the cochlea–transforms the acoustical spectrum into a spatial tonotopical map [12]. Cochlear masking is the phenomenon whereby the 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 a logarithmic scale where the spacing and the masker frequency slopes extend more regularly to either side of the spectrum [13]. 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.

Masking is investigated 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 $L_m = 60\,\text{dB}$ (see Figure 6.14 of [14]). 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.

The basic model for simultaneous masking consists of a linear masker threshold, with slopes of $+30\,\text{dB}$ per band for lower band masking and $-8\,\text{dB}$ per band for higher band masking [14].

Temporal masking has methodologically been treated as two separate processes: *premasking* occurs before the appear-



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

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

(c) Final cochleogram S'.

Fig. 2: Choice spectrograms output by each step of the architecture.

ance of the masker while *postmasking* manifests itself after the masker is no longer present. It is generally 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 [11].

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 [15],

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

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.

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 [14]. 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.

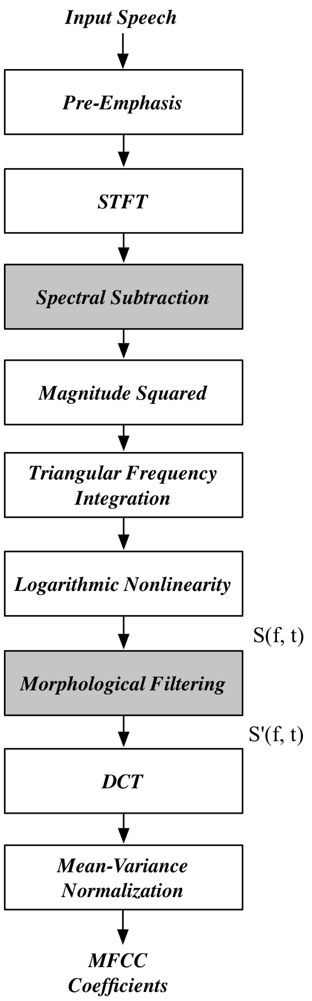


Fig. 3: Structure of the proposed front-end algorithm; shaded blocks, spectral subtraction (SS) and Morphological Filtering (MF), indicates the major differences regarding the conventional MFCC algorithm.

Our SE is 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 1.

Different sizes in both frequency and time scale were tested in [1], 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 was applied. Finally, the SE was padded with zeros around the pixel in which the morphological closing operation is to be performed. The SE can be seen at the upper left of Figure 2(a) along with examples of the output of some of the processing steps leading to the final cochleogram.

C. 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 traditional MFCC spectro-temporal representation. Figure 3 represents the block diagram of the complete proposed front-end based on Mel-frequency spectro-temporal representations where the gray blocks are our additions to the conventional MFCC 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 the single structuring element described in the previous section. As for the spectral subtraction block, we found synergies with MF under the MFCC framework in previous works [11], [16], [17] that we also confirm in this paper for a deep architecture. The last two blocks 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).

III. Deep neural networks and hybrid speech recognition systems

A Deep Neural Network (DNN) is a Multi-Layer Perceptron (MLP) with a larger number of hidden layers between its inputs and outputs, whose weights are fully connected and are often initialized using an unsupervised pre-training scheme.

As a traditional MLP, the feed-forward architecture can be computed as follows:

$$\mathbf{h}^{(l+1)} = \sigma\left(\mathbf{W}^{(l)}\mathbf{h}^{(l)} + \mathbf{b}^{(l)}\right), \quad 1 \le l \le L \qquad (4)$$

where $\mathbf{h}^{(l+1)}$ is the vector of inputs to the l+1-th layer, $\sigma(x) = (1+e^{-x})^{-1}$ is the sigmoid activation function, L is the total number of hidden layers, $\mathbf{h}^{(l)}$ is the output vector of hidden layer l and $\mathbf{W}^{(l)}$ and $\mathbf{b}^{(l)}$ are the weight matrix and bias vector of layer l, respectively.

Training a DNN using the well-known error back-propagation (BP) algorithm with a random initialization of its weight matrices may not provide a good performance as it may become stuck in a local minimum. To overcome this problem, DNN parameters are often initialized using an unsupervised technique, e.g. Restricted Bolzmann Machines (RBMs) [18] or Stacked Denoising Autoencoders (SDAs) [19].

A. Hybrid Speech Recognition Systems

In a hybrid DNN/HMM system, just as in classical ANN/HMM hybrid architectures [20], a DNN is trained to classify the input acoustic features into classes corresponding to the states of HMMs in such a way that the state emission likelihoods usually computed with GMM are replaced by the likelihoods generated by the DNN.

The DNN estimates the posterior probability $p(s|\mathbf{o}_t)$ of each state s given the observation $\mathbf{o}_t$ at time t through a softmax final layer

$$p(s|\mathbf{o}_t) = \frac{\exp\left(\mathbf{W}^{(L)}\mathbf{h}^{(L)} + \mathbf{b}^{(L)}\right)}{\sum_{\bar{S}} \exp\left(\mathbf{W}^{(L)}\mathbf{h}^{(L)} + \mathbf{b}^{(L)}\right)}\,. \qquad (5)$$

In a hybrid ASR system, the HMM topology is set from a previously trained GMM-HMM, and the DNN training data come from the forced-alignment between the state-level transcripts and the corresponding speech signals obtained by using this initial GMM-HMM system. In the recognition stage, the DNN estimates the emission probability of each HMM state. Bayes' rule is used to obtain the state emission likelihoods $p(\mathbf{o}_t|s)$ and the $p(s|\mathbf{o}_t)$ estimated by the DNN is scaled by the class prior, $p(s)$, which can be estimated by counting the occurrences of each state on the training data.

B. Dropout

The most important problem to overcome in DNN training is overfitting. This problem usually arises when we train a large DNN with a small training set. A training method called *dropout* proposed in [6] tries to reduce overfitting and improves the generalization capability of the network by randomly omitting a certain percentage of the hidden units on each training iteration.

When dropout is employed, the activation function of (4) can be rewritten as:

$$\mathbf{h}^{(l+1)} = m^{(l)} \star \sigma\left(\mathbf{W}^{(l)}\mathbf{h}^{(l)} + \mathbf{b}^{(l)}\right), \quad 1 \le l \le L \qquad (6)$$

where $\star$ denotes the Hadamard (entry-wise) product, and $m^{(l)}$ is a binary vector conformal to $\mathbf{h}^{(l)}$ whose elements are sampled from a Bernoulli distribution with probability p. This probability is the so called *Hidden Drop Factor* (HDF) and must be determined over a validation set as explained in Section IV.

As the *sigmoid* function has the property that $\sigma(0) = 0$, Eq. (6) can be rewritten as:

$$\mathbf{h}^{(l+1)} = \sigma\left(m^{(l)} \star \left(\mathbf{W}^{(l)}\mathbf{h}^{(l)} + \mathbf{b}^{(l)}\right)\right), \quad 1 \le l \le L \qquad (7)$$

where dropout is applied on the inputs of the activation function, leading to a more efficient way of performing dropout training. Note that dropout is only applied in the training stage whereas during testing all the hidden units become active.

Dropout networks are trained with the standard stochastic gradient descent algorithm but using the forward architecture presented on Eq. (6) instead of Eq. (4). Following [21], during testing we compensate the parameters to take into account the dropout factor by scaling the weight matrices as follows:

$$\overline{\mathbf{W}}^{(l)} \;=\; (1 - HDF) \cdot \mathbf{W}^{(l)} \qquad (8)$$

Dropout has already successfully tested on noise robust ASR in [22]. Its benefits come from the improved generalization abilities attained by reducing their capacity. Another interpretation of the behaviour of dropout is that in the training stage it adds random noise to the training set resulting in a network that is very robust to variabilities in the inputs (in our particular case, due to the addition of noise).

C. Maxout Deep Neural Network

A Maxout Deep Neural Network (DMN) [7] is a modification of the feed-forward architecture (Eq. (4)) where the maxout activation function is employed. The maxout unit simply takes the maximum over a set of inputs. In a DMN each hidden unit takes the maximum value over the g units of

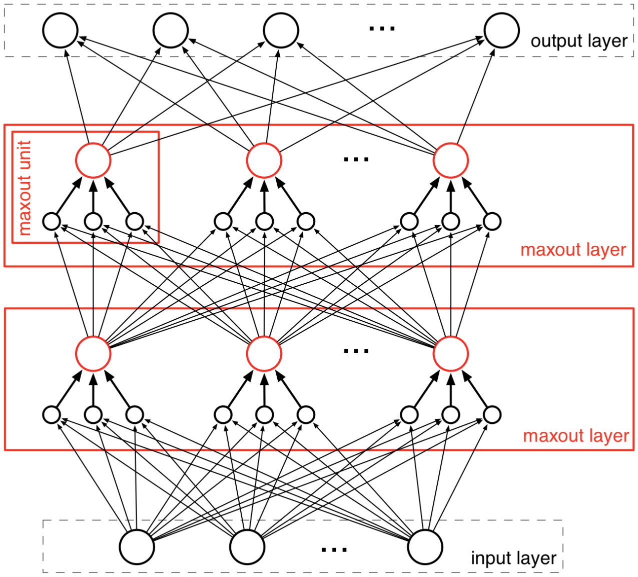


Fig. 4: A Maxout Network of 2 hidden layers and a group size of $g = 3$. The hidden nodes in red perform the max operation.

a group. The output of the hidden node i of the layer $l+1$ can be computed as follows:

$$h_i^{(l+1)} = \max_{j \in 1,\dots,g} z_{ij}^{(l+1)}, \quad 1 \le l \le L \tag{9}$$

where $z_{ij}^{(l+1)}$ are the linear pre-activation values from the l layer:

$$\mathbf{z}^{(l+1)} = \mathbf{W}^{(l)}\mathbf{h}^{(l)} + \mathbf{b}^{(l)} \tag{10}$$

As can be observed the max-pooling operation is applied over the $\mathbf{z}^{(l+1)}$ vector. Note that DMNs fairly reduce the number of parameters over DNNs, as the weight matrix $\mathbf{W}^{(l)}$ of each layer in the DMN is $1/g$ of the size of its equivalent DNN weight matrix. This makes DMN more convenient for ASR tasks where the training sets and the input and output dimensions are normally very large. An illustration of a DMN with 2 hidden layers and a group size of $g = 3$ is shown in Figure 4.

In [7] a demonstration of the capability of maxout units to approximate any convex function by tuning the weights of the previous layers is included. For this purpose, the shapes of activation functions are not fixed, allowing the DMNs to model the variability of speech more smoothly. DMNs are commonly applied in conjunction with dropout to maximize the model averaging effects of dropout

IV. Experiments

Our experiments for evaluating and comparing the performance of conventional GMM-HMM and the different hybrid deep neural networks-based ASR systems with our biologically inspired features on the TIMIT corpus [23] are presented below. In particular, we used the 462 speaker training set, a development set of 50 speakers to tune all the parameters and finally the 24 speaker core test set. Each utterance is recorded at 16 kHz and the corpus includes time-aligned phonetic transcriptions allowing us to give results in terms of Phone Error Rate (PER).

To test the robustness of the different methods we added four different types of noises (white, street, music and speaker) at four different SNRs using the FANT tool [24] (with G.712 filtering) to the clean speech database. These noises are the ones used in [25]. All the noisy tests are evaluated in mismatch conditions (that is, training with clean conditions and testing on noisy speech).

We used the Kaldi toolkit [26] for implementing the traditional GMM-HMM ASR system and the PDNN toolkit [27] for the hybrid DNN-based ASR systems.

In all the cases, the input features were 12th-order modified MFCCs plus a log-energy coefficient, and their corresponding first and second order derivatives yielding a 39 component feature vector. Mean and variance normalization on each of the components was applied. A context of 5 frames was chosen for the hybrid models. All the hybrid systems were trained with the labels generated from the best performance GMM-HMM system through forced alignment.

We chose as baseline the Deep Maxout Networks (DMNs) in combination with dropout, since we prove in [8] that DMNs perform better in almost every situation for all the noises considered in comparison to other systems (Monophone, Triphone, Triphone with Lineal Discriminant Analysis, Maximum Likelihood Linear Transform, and Speaker Adaptative Training, traditional DNNs with and without pretrain and DNNs with dropout).

The configuration parameters of the network (number of hidden layers, HDF, group size and momentum when applicable) are set up based on previous work [8] where: HDF is 0.2 and the group size $g = 3$. The number of hidden maxout units for the DMN is 400.

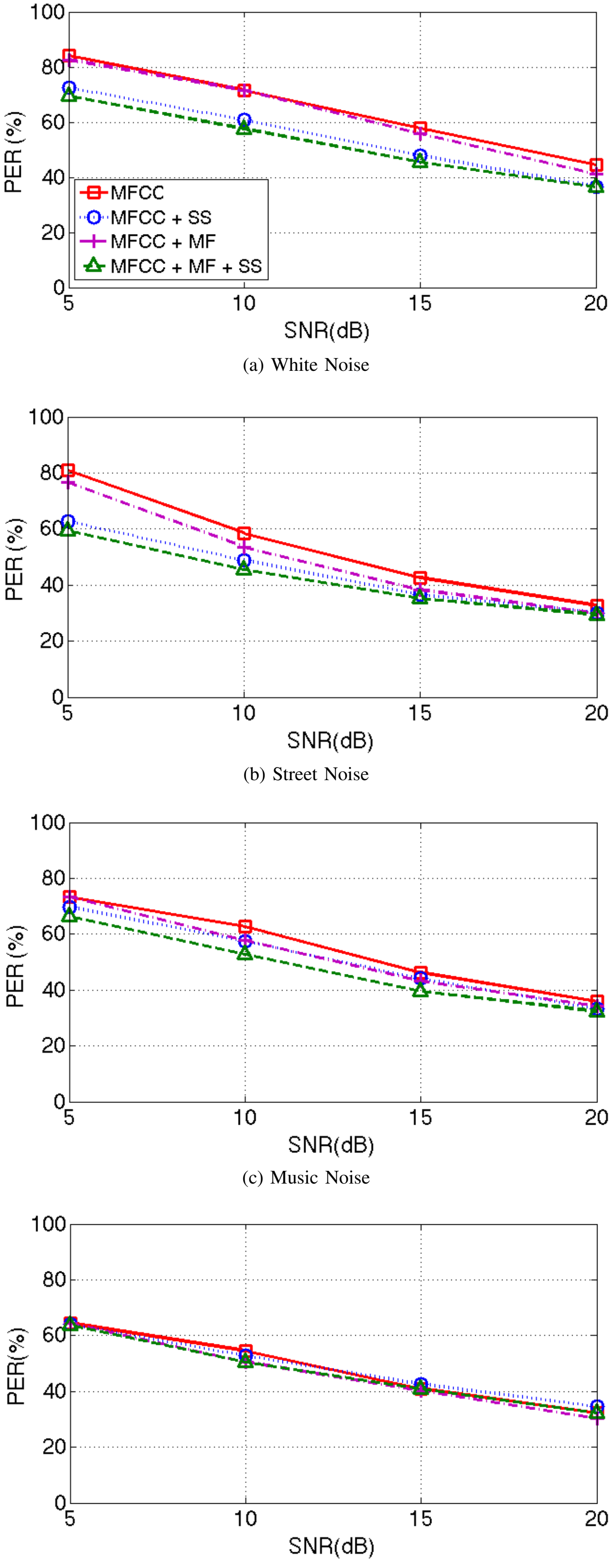
Figure 5 shows the Phone Error Rate (PER) for each type of noise and SNR, obtained by the baseline (MFCC), the baseline with spectral subtraction (SS), and our features with spectral subtraction (MF + SS) and without spectral subtraction (MF).

Figure 5 shows that: (1) the application of MF improves the baseline recognition rates for all noises except for speaker noise where all the results are very similar; (2) the SS technique also improves the baseline in all cases; (3) the joint use of SS and MF improves the recognition rates obtained with SS and with the baseline, and; (4) the (MF+SS) method achieves the best performance in almost every noise and SNR conditions.

V. Conclusions and Future Work

In this paper the combination of biologically motivated features and Deep Maxout Networks (DMNs) is employed for robust speech recognition into a hybrid DNN-HMM ASR system, showing a better performance than conventional MFCC and spectral subtraction on the same architecture. The proposed features are designed taking into account the HAS masking properties through the proper choice of the SE and the application of morphology operations on the cochleograms.

Results show that the application of morphological processing in conjunction with spectral subtraction produces a



(a) White Noise

(b) Street Noise

(c) Music Noise

(d) Speaker Noise

Fig. 5: Comparison of the performance of the different systems in terms of PER [%] for TIMIT test set in different noisy conditions.

significant increase in recognition rates on a noisy version of the Timit dataset. Also, it is proved that the DMN-based back-end is capable of take advantage of these auditorily inspired features making the whole system more robust, suggesting that research in new acoustic representations of speech still has an important role in the DNN-based automatic speech recognition systems.

Future work will focus on two directions. Regarding the feature extraction process, we plan to introduce the dependency of the masker strength into the morphological procedure. With respect to the back-end, further lines of research include testing the DMN in bigger datasets and with other novel machine learning techniques, like *dropconnect* [28].

VI. Acknowledgements.

This contribution has been supported by an Airbus Defense and Space Grant (Open Innovation - SAVIER) and Spanish Government-CICYT project TEC2014-53390-P. We would also like to thank Chanwoo Kim for kindly providing the testing noises.

References

[1] 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 *Proceedings of Interspeech (15th International Conference on Speech Communication and Technology)*, 2014, pp. 2430 – 2434. 1, 3

[2] G. E. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, 2012. 1

[3] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 20, no. 1, 2012. 1

[4] N. Morgan, "Deep and wide: Multiple layers in automatic speech recognition," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 20, no. 1, 2012. 1

[5] A. Mohamed, G. E. Dahl, and G. E. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 20, no. 1, 2012. 1

[6] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *CoRR*, 2012. 1, 4

[7] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout Networks," *ArXiv e-prints*, Feb. 2013. 1, 4, 5

[8] F. de-la Calle-Silos, A. Gallardo-Antolín, and C. Peláez-Moreno, "Deep Maxout Networks applied to noise-robust speech recognition," in *Advances in Speech and Language Technologies for Iberian Languages*, ser. Lecture Notes in Computer Science, J. L. Navarro Mesa, A. Ortega, A. Teixeira, E. Hernández Pérez, P. Quintana Morales, A. Ravelo García, I. Guerra Moreno, and D. T. Toledano, Eds. Springer, 2014, vol. 8854, pp. 109–118. 1, 5

[9] G. Matheron and J. Serra, "The birth of mathematical morphology," in *Proc. 6th Int. Symp. Mathematical Morphology*. Sydney, Australia, 2002, pp. 1–16. 2

[10] E. R. Dougherty and R. A. Lotufo, *Hands-on Morphological Image Processing*, ser. Tutorial Texts in Optical Engineering. SPIE press, 2003. 2

[11] 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. 2, 3, 4

[12] 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. 2

[13] 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. 2

[14] H. Fastl and E. Zwicker, *Psycho-acoustics: Facts and Models*, 3rd ed. Springer, 2007. 2, 3

[15] 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. 3

[16] 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. 4

[17] J. Cadore, C. Peláez-Moreno, and A. Gallardo-Antolín, “Morphological processing of a dynamic compressive gammachirp filterbank for automatic speech recognition,” in *IberSPEECH 2012*, 2011. 4

[18] G. E. Hinton, “A practical guide to training restricted boltzmann machines.” ser. Lecture Notes in Computer Science, G. Montavon, G. B. Orr, and K.-R. Mller, Eds. Springer, 2012, vol. 7700. 4

[19] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, “Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion,” *Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, Dec. 2010. 4

[20] H. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*, ser. Kluwer international series in engineering and computer science: VLSI, computer architecture, and digital signal processing. Springer US, 1994. 4

[21] Y. Miao and F. Metze, “Improving low-resource CD-DNN-HMM using dropout and multilingual DNN training.” in *INTERSPEECH*. ISCA, 2013, pp. 2237–2241. 4

[22] M. L. Seltzer, D. Yu, and Y. Wang, “An investigation of deep neural networks for noise robust speech recognition,” in *Acoustics, Speech, and Signal Processing (ICASSP), IEEE International Conference on,*, 2013. 4

[23] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, “DARPA TIMIT acoustic phonetic continuous speech corpus cdrom,” https://catalog.ldc.upenn.edu/LDC93S1, 1993. 5

[24] G. Hirsch, “Fant - filtering and noise adding tool,” http://dnt.kr.hsnr.de/download.html, 2005. 5

[25] C. Kim and R. M. Stern, “Power-normalized cepstral coefficients (PNCC) for robust speech recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*. 5

[26] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, “The Kaldi speech recognition toolkit,” in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011. 5

[27] Y. Miao, “Kaldi+PDNN: Building DNN-based ASR systems with Kaldi and PDNN,” *CoRR*, 2014. 5

[28] L. Wan, M. D. Zeiler, S. Zhang, Y. LeCun, and R. Fergus, “Regularization of neural networks using dropconnect,” in *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, 2013. 6