Interspeech 2018
2-6 September 2018, Hyderabad

Biophysically-inspired features improve the generalizability of neural network-based speech enhancement systems

Deepak Baby, Sarah Verhulst

Dept. of Information Technology, Ghent University, Belgium
{deepak.baby, s.verhulst}@ugent.be

Abstract

Recent advances in neural network (NN)-based speech enhancement schemes are shown to outperform most conventional techniques. However, the performance of such systems in adverse listening conditions such as negative signal-to-noise ratios and unseen noises is still far from that of humans. Motivated by the remarkable performance of humans under these challenging conditions, this paper investigates whether biophysically-inspired features can mitigate the poor generalization capabilities of NN-based speech enhancement systems. We make use of features derived from several human auditory periphery models for training a speech enhancement system that employs long short-term memory (LSTM), and evaluate them on a variety of mismatched testing conditions. The results reveal that biophysically-inspired auditory models such as nonlinear transmission line models improve the generalizability of LSTM-based noise suppression systems in terms of various objective quality measures, suggesting that such features lead to robust speech representations that are less sensitive to the noise type.
Index Terms: speech enhancement, neural networks, auditory models, long short-term memory

1. Introduction

Speech signals captured from realistic acoustic scenarios are typically very complex with added degradations such as background noise and reverberation. The goal of speech enhancement systems is to improve the intelligibility and quality of such degraded speech signals. Since the performance of several applications such as automatic speech recognition, mobile communication and hearing aids depends on the quality of the captured audio signal, speech enhancement has been an actively researched topic and several methods have been proposed over the past several decades [1, 2].

Recent advances in deep neural network (DNN)-based learning architectures are shown to outperform most of the conventional speech enhancement approaches [3–7], thanks to their nonlinear structure with multiple hidden layers which enables them to model the complex degradations in the captured speech signal. However, most DNN-based speech enhancement studies limit the evaluation to *matched testing conditions* where the noise types and level (signal-to-noise ratio, SNR) of the test data are similar to that of the training data. Since DNN approaches are data-driven, it is indeed expected that they will outperform the conventional signal processing-based methods in matched noise conditions when enough training data is available [8]. Therefore, a thorough analysis on the generalizability of DNN-based speech enhancement to *mismatched* test conditions needs to be done.

While there exist techniques such as dropout [9] and weight regularization [10] to reduce overfitting to the training data, such approaches have limited applicability when the input features are heavily corrupted from different noise conditions. There exist a few studies that have attempted to improve the generalizability of DNN-based systems in mismatched noise conditions. In one approach [4], a DNN is trained using a large variety of noise types and they show that significant improvements can be achieved in mismatched test conditions. However, this approach still requires a lot of training noise examples and some of the test conditions were similar to those present in the training set (e.g., the car and exhibition noise conditions in the test data would be similar to a few training noise conditions such as Traffic and Car Noise, and Crowd Noise.). This work concentrates on investigating the generalizability of DNN-based systems when only a few noise conditions are available during training.

Although several core concepts of DNNs stem from the cortical processing in the human brain [11], most DNN-based speech enhancement systems still make use of conventional representations of the acoustic speech signal such as short-time Fourier transform (STFT) or Mel-integrated magnitude STFT (dubbed *FBANK* features) [4, 6–8, 12]. Since humans perceive speech remarkably well under a large variety of adverse listening conditions [13], we argue that biophysically inspired representations of speech might lead to a more generalizable DNN system. The idea of using auditory inspired features for DNN-based applications itself is not new. It has been previously shown that auditory inspired features such as modulation spectrogram [14] and Gabor filter-bank [15] features improve the performance of speech recognition systems employing fully-connected DNNs.

This paper further develops this work by combining the latest advances in human auditory modeling with state-of-the-art DNN-based speech enhancement systems. For this, we make use of biophysically inspired models of the human cochlea such as nonlinear transmission line models [16] and with recurrent neural networks (RNNs) for speech enhancement. In a previous study, we showed that biophysically-inspired features improve the generalizability of fully-connected NN-based speech enhancement systems [17] with a small training dataset ($\approx$ 1hr). Here, we extend this approach by using RNNs which employ long short-term memory (LSTM-RNNs) cells intending to leverage upon its memory structure that can capture temporal contexts with a larger training data.

This work investigates several LSTM-RNN-based speech enhancement systems that are trained using different cochlear models and we systematically evaluate and compare them under a variety of mismatched test conditions with differing noise types and SNR levels. The main contributions of this work are: 1) combine the best of two advanced research fields, viz. auditory models and deep learning, and 2) investigate whether such a combination yields better generalizable systems, even in conjunction with complex LSTM-based models.

10.21437/Interspeech.2018-1237

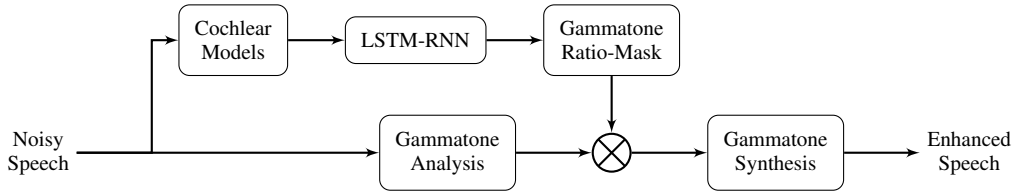


Figure 1: *Block diagram overview of the investigated neural network-based speech enhancement system.*

2. Speech enhancement using LSTM-RNNs

The goal of a single-channel speech enhancement system is to recover the underlying clean speech signal $s[n]$ from the noisy speech recording $y[n] = s[n] + w[n]$, where $w[n]$ is the added noise. In the spectro-temporal domain, let $\mathbf{Y}(b, f)$, $\mathbf{S}(b, f)$ and $\mathbf{W}(b, f)$ be the time-frequency representations of $y[n]$, $s[n]$ and $w[n]$ respectively at frequency-bin/band index b and frame index f. We make use of the Gammatone filter-bank (comprised of B bands) energies computed over windows to represent the noisy speech signal in the spectro-temporal domain.

The LSTM-RNNs are trained to predict the ideal ratio-mask (IRM) for enhancing the noisy Gammatone spectrogram as it is shown to perform better than the ideal binary masks [18]. The IRM is defined as:

$$\text{IRM}(b, f) \triangleq \frac{\mathbf{S}(b, f)}{\mathbf{S}(b, f) + \mathbf{W}(b, f)}. \quad (1)$$

The noisy Gammatone spectrogram $\mathbf{Y}$ is enhanced by computing $\mathbf{Y}(b, f) \cdot \text{IRM}(b, f)$ after which Gammatone synthesis [12] is applied to reconstruct the enhanced speech signal in the time-domain. A block diagram overview of the procedure is depicted in Figure 1.

The features derived from various auditory models are fed as input to the LSTM-RNN system that predicts the IRM in the Gammatone domain. The final LSTM layer uses sigmoid activation at its output. The network weights are trained using the back-propagation through time (BPTT) algorithm such that the mean-square error between the predicted and the target IRMs is minimized. In essence, this LSTM configuration operates as a seq-to-seq system that generates IRM sequences corresponding to the input feature sequence.

2.1. Auditory inspired features

The input features for the LSTM-RNN-based speech enhancement system are derived from various auditory models using a two-step approach. First, the audio signal is segregated into B_c cochlear channels using the auditory models. The output of this stage is an activity pattern along the cochlea in response to the input audio signal. This representation will be of size $B_c \times L$, where L is the length of the time domain signal. In the second step, the temporal dimension of the activity pattern is reduced by computing the energy over the same window-length as used for the Gammatone energies, since the output targets are Gammatone IRMs. This results in input features and targets with the same temporal dimension F. Thus, the input features are of size $B_c \times F$ and the target IRMs for enhancing Gammatone energies are of size $B \times F$.

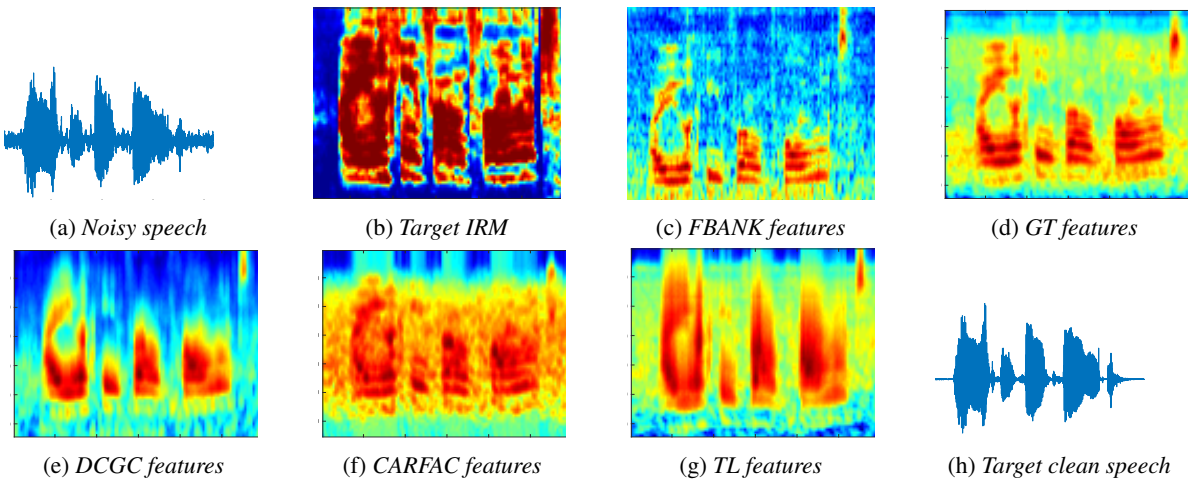
The various auditory models investigated in this study are:
*1. Gammatone filter-bank (**GT**)*: This filter-bank consists of a set of parallel filters that approximate the shape, sharpness and bandwidth of human auditory filters [19]. Auditory filtering is attributed to the mechanics of the cochlea that result in a set of bandpass filters with decreasing center frequency and increasing sharpness as sound travels from the cochlear base (close to the middle ear) to the apex. GT filters can easily be inverted, which motivates their use for the enhancement phase of our processing (Figure 1).
*2. Dynamically compressed Gammachirp (**DCGC**)*: This filter-bank incorporates realistic level-dependent changes in auditory filter tuning, that yield wider filters for higher sound pressure levels [20]. In a nutshell, the DCGC model consists of a set of parallel GT filters that are followed by a level-dependent high-pass asymmetric function that mimics the active and compressive action of cochlear outer-hair-cells, which are responsible for the level-dependent tuning in humans.
*3. Cascade of asymmetric resonators with fast acting compression (**CARFAC**)*: Where the previous models showed a parallel filter-bank architecture, CARFAC follows a serial (i.e., cascaded) approach in which each filter output serves as the input to the next [21]. This architecture is more similar to the mechanical structure of the cochlea in which sound travels from the base to the apex over the longitudinally coupled basilar-membrane. CARFAC is implemented as a cascade of second-order filters with one complex conjugate pair of zeros and poles. The individual filters act as second-order asymmetric resonators, whose characteristics (i.e., best frequency and level-dependent damping ratio) are set to match human cochlear filter tuning.
*4. Nonlinear transmission-line model (**TL**)*: This model approximates the cochlear processing as a cascade of shunt admittances and serial impedances that model the mechanical filter properties and fluid coupling in the cochlea, respectively. The model parameters are set to yield realistic human cochlear filter bandwidths that vary as a function of frequency and level [16, 22, 23]. The cascaded over parallel organization of the bandpass filters results in capturing cochlear phenomena related to coupling: e.g. two-tone suppression, frequency glides, traveling waves.

The various input features which are derived from these models together with the target IRM for a noisy speech signal (Babble noise added at 3dB SNR) are depicted in Figure 2.

3. Evaluation setup

To evaluate the speech enhancement system under different training and mismatched testing conditions, recordings from the TIMIT dataset (16kHz sampling frequency) were used to generate the various noisy datasets. We investigated three different noise conditions: Babble, ICRA and Factory noises. ICRA is a non-stationary noise designed for clinical testing of hearing aids [24] with spectral and temporal characteristics similar to real-life speech and babble noise. The babble and factory noise recordings were taken from the NTT Ambient noise database.

From the 3696 utterances in the TIMIT training dataset, recordings longer than 5s were omitted since the maximum in-



(a) *Noisy speech* (b) *Target IRM* (c) *FBANK features* (d) *GT features*

(e) *DCGC features* (f) *CARFAC features* (g) *TL features* (h) *Target clean speech*

Figure 2: *The various input feature representations ((c) - (g)) together with the target IRM (b) corresponding to a noisy speech signal (a) containing Babble noise at* 3 *dB SNR. The logarithm of the values are plotted for a better visualization. The horizontal and the vertical axes correspond to the time and frequency axes, respectively.*

put sequence length of the LSTM-RNN used in this study was 5 seconds. From the resulting 3 577 utterances, two training sets were generated, each containing Babble and ICRA noises that were added at a random SNR between 6 and 12 dB. We thus trained 10 different LSTM-based speech enhancement systems (2 training sets $\times$ 5 cochlear models). The test set was comprised of 9 noise conditions (3 noise types $\times$ 3 SNRs) in which Babble, ICRA and Factory noises were added at -3, 3 and 9dB SNR levels. The core test of the TIMIT database containing 192 recordings was also pruned to have a maximum length of 5 seconds and the resulting 187 utterances were used to create the test datasets.

The Gammatone features were extracted using the implementation provided with the auditory modeling toolbox [25]. The DCGC and CARFAC representations were obtained using the auditory image modeling toolbox (AIM-MAT) [26] and the implementation provided in [27], respectively. The implementation for the TL model was obtained from [28]. All these models were set to use $B = 64$ cochlear channels and the envelope energies were computed over a window-length of 20 ms shifted by 10 ms, resulting in 100 feature vectors per second. The logarithm of the energies together with their Δ coefficients were used as input to the LSTM as they were found to yield better results. The LSTMs were trained to predict the IRMs corresponding to the 64 gammatone bands. The input and output feature dimensions of the trained LSTM-RNN were thus 128 and 64, respectively.

The LSTM-RNN setting was comprised of two hidden LSTM layers containing 512 cells each and an output LSTM layer with 64 cells and sigmoid activation function that generate the IRMs. As mentioned before, the maximum input sequence length was set as 5 seconds, i.e., 500 frames. The input and output features that were less than 500 frames were zero-padded to match with the maximum sequence length. The training and test features were mean and variance normalized using the mean and variance of the training set. A batch-size of 16 utterances was used to train the LSTM-RNN using the Adam optimizer with a learning rate of 0.0001 for 200 epochs. Dropout with a keep probability of 0.8 was also used at every LSTM layer to reduce overfitting. A validation set containing the same noise in the training data at 3dB SNR was used for parameter tuning. The LSTM-RNN models were defined and trained using the Tensorflow toolkit [29] and GPUs were used to accelerate the training using BPTT.

During the test phase, the IRMs generated from the trained LSTM-RNNs were used to enhance the Gammatone energies, and Gammatone synthesis [12] was used to reconstruct the enhanced time-domain speech signal. To evaluate and compare the various speech enhancement systems, the following objective measures are used: perceptual evaluation of speech quality (PESQ) in terms of mean opinion score (MOS), segmental SNR (segSNR) and cepstral distance (CD) in dB. Higher values of PESQ and SRMR, and lower values of CD indicate a better performance. For better readability, the improvements in these measures (denoted as ΔPESQ, ΔsegSNR and ΔCD) are used for comparing the results. The Δs for the PESQ and segSNR were obtained by subtracting the metric obtained on the noisy data from that of the enhanced data, whereas the ΔCD measure is obtained by subtracting the metric obtained on the enhanced data from that of the noisy data. In short, a higher Δ value implies a better performance for all the measures.

4. Results and discussion

The noise suppression performance evaluated using various speech quality measures on different test conditions are provided in Table 1. We also include a conventional spectral subtraction-based speech enhancement system (shown as SS) described in [30] as an additional baseline system. It can be seen that the SS approach performs worse when compared to the supervised LSTM-based settings as the considered noise types are non-stationary.

As expected, all models yielded a similar performance in presence of matched noise conditions. The biophysically-inspired nonlinear models, especially the cascaded TL and CARFAC models, resulted in a more generalizable DNN system as they consistently showed a better performance in mismatched noise conditions. The DCGC model outperformed the GT model only in a few conditions and the improvements were not consistent across different training and test conditions. This could be attributed to the dynamic compression in the DCGC that behaves differently for different noise conditions. It can

Table 1: *Comparison of noise suppression performance of the various LSTM-based speech enhancement systems trained using the various auditory models. For all the metrics, a higher value means a better performance.*

	Training Data: Bab 6-12 dB									***Training Data: ICRA 6-12 dB***								
	Babble			**ICRA**			**Factory**			**Babble**			**ICRA**			**Factory**		
	-3dB	**6dB**	**9dB**	**-3dB**	**6dB**	**9dB**	**-3dB**	**6dB**	**9dB**	**-3dB**	**6dB**	**9dB**	**-3dB**	**6dB**	**9dB**	**-3dB**	**6dB**	**9dB**
	ΔPESQ																	
FBANK	0.17	0.51	0.78	0.07	0.20	0.40	0.08	0.29	0.46	0.07	0.28	0.50	0.26	0.57	0.88	0.07	0.21	0.42
GT	0.17	0.51	0.78	0.05	0.19	0.33	0.08	0.27	0.44	0.07	0.29	0.50	0.26	0.57	0.86	0.06	0.22	0.39
DCGC	0.17	0.48	0.74	0.02	0.17	0.38	0.07	0.26	0.43	0.04	0.32	0.52	0.26	0.53	0.87	0.06	0.23	0.41
CARFAC	0.15	0.48	0.74	0.08	0.23	0.42	0.11	0.30	0.48	0.11	0.35	0.60	0.23	0.57	0.88	0.13	0.30	0.42
TL	0.16	0.51	0.77	0.11	0.26	0.47	0.15	0.32	0.52	0.13	0.38	0.68	0.25	0.60	0.90	0.18	0.35	0.50
SS	0.05	0.19	0.28	0.03	0.15	0.24	0.07	0.17	0.31	0.05	0.19	0.28	0.03	0.15	0.24	0.07	0.17	0.28
	ΔsegSNR																	
FBANK	6.87	8.47	7.50	2.89	4.08	3.10	4.21	6.03	5.34	4.26	5.98	5.66	8.78	9.82	8.15	3.38	5.29	5.01
GT	7.07	8.55	7.49	2.06	3.10	2.51	3.95	5.63	5.21	4.42	6.22	5.81	9.24	10.38	8.45	3.12	5.14	4.85
DCGC	6.60	8.00	7.11	2.19	3.10	2.64	3.99	5.36	5.01	4.49	6.27	5.92	9.09	10.11	8.14	3.23	5.18	4.95
CARFAC	6.84	8.40	7.37	3.24	4.45	3.57	4.65	6.33	5.62	5.03	6.65	6.28	9.08	10.14	8.15	3.59	5.87	5.11
TL	6.92	8.48	7.46	3.42	4.59	3.69	4.95	6.44	5.73	5.14	6.86	6.37	9.12	10.29	8.36	3.89	6.08	5.56
SS	1.43	2.79	2.94	1.03	2.21	2.54	1.94	3.27	3.04	1.43	2.79	2.94	1.03	2.21	2.54	1.94	3.27	3.04
	ΔCD																	
FBANK	1.27	1.64	1.53	0.40	0.80	0.85	0.88	1.21	1.28	0.76	1.04	1.06	1.68	1.86	1.71	0.68	1.00	1.10
GT	1.26	1.63	1.52	0.37	0.60	0.66	0.81	1.17	1.18	0.70	1.01	1.05	1.67	1.91	1.72	0.63	0.96	0.98
DCGC	1.14	1.52	1.44	0.37	0.59	0.68	0.83	1.17	1.16	0.73	1.09	1.11	1.62	1.84	1.67	0.65	0.96	0.99
CARFAC	1.16	1.51	1.47	0.44	0.81	0.85	0.91	1.23	1.30	0.86	1.24	1.24	1.61	1.83	1.66	0.72	1.06	1.12
TL	1.21	1.59	1.50	0.50	0.89	0.92	0.98	1.30	1.35	0.91	1.31	1.28	1.64	1.87	1.68	0.80	1.15	1.21
SS	0.24	0.37	0.31	0.19	0.31	0.28	0.31	0.43	0.39	0.24	0.37	0.31	0.19	0.31	0.28	0.31	0.43	0.39

also be seen from Figure 2 that the dynamic compression seems to distort the high frequency regions especially in presence of noise. The FBANK features were observed to generalize well when trained on babble noise, but this was not observed for the ICRA noise training condition. Another interesting observation is that the FBANK features yielded good results under high SNR conditions and that the performance dropped considerably at lower SNRs suggesting that FBANK features are sensitive to the mismatch in SNR levels as well.

In general, it can be seen that the cascaded models such as CARFAC and TL lead to a better generalizable DNN system when compared to the parallel filter-bank models. This shows the benefits and potential of using biophysically inspired, cascaded filtering models of the cochlea for speech related applications. The reason why cascade filter-models perform better than their parallel counterparts can be explained by the SNR improvement that is obtained when considering a single filter in the cascade. It was previously shown that the longitudinal coupling of filters (even if they have the same tuning) results in a 2–5 dB SNR improvement at the filter output, for tone-in-noise stimuli [31]. This shows that if we capture the complexity of cochlear mechanics (yielding a natural noise-reduction) in the features provided to the DNN, then such systems themselves become more generalizable and robust to different testing conditions. Additionally, describing the cochlear mechanics realistically requires more computational effort than when filtering using a GT filterbank, which might pose a design constraint on the desired application.

5. Conclusions

This paper investigated whether biophysically inspired features can mitigate the poor generalization capabilities of an LSTM-RNN-based speech enhancement system. We compared 5 different cochlear models ranging from traditional FBANK features to the latest transmission line model. The evaluations under a variety of matched and mismatched test conditions revealed that the biophysical models such as TL and CARFAC models improve the generalizability of LSTM-RNN-based speech enhancement systems. It was also observed that simple nonlinear models such as DCGC did not result in a better speech enhancement setting. The promising results with TL and CARFAC models also highlights the mutual benefits of combining the latest advances in human auditory modeling and DNNs.

Investigating the robustness of such features for state-of-the-art speech recognition systems is a suggested future work. Since these models are also capable of modeling higher levels of the human auditory pathway such as auditory nerve fibre responses, another promising research direction would be to investigate features derived from such higher level speech representations as well.

6. Acknowledgements

This work has been funded with support from the European Union's Horizon 2020 research and innovation programme under grant agreement No 678120 (RobSpear). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

7. References

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. Boca Raton, FL, USA: CRC Press, Inc., 2013.

[2] I. Cohen and S. Gannot, *Spectral Enhancement Methods*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 873–902.

[3] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Au-*

dio, Speech, and Language Processing, vol. 22, no. 12, pp. 1849–1858, Dec 2014.

[4] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech & Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[5] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. INTERSPEECH*. Lyon, France: ISCA, August 2013, pp. 436–440.

[6] L. Sun, J. Du, L. Dai, and C. Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," in *Hands-free Speech Communications and Microphone Arrays, HSCMA*, San Francisco, CA, USA, March 2017, pp. 136–140.

[7] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Deep recurrent networks for separation and recognition of single-channel speech in nonstationary background audio," in *New Era for Robust Speech Recognition, Exploiting Deep Learning.*, 2017, pp. 165–186.

[8] A. Kumar and D. A. F. Florêncio, "Speech enhancement in multiple-noise conditions using deep neural networks," in *Proc. INTERSPEECH*. San Francisco, CA, USA: ISCA, September 2016, pp. 3738–3742.

[9] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[10] J. Z. Kolter and A. Y. Ng, "Regularization and feature selection in least-squares temporal difference learning," in *Proceedings of the 26th Annual International Conference on Machine Learning ICML*, Montreal, Quebec, Canada, June 2009, pp. 521–528.

[11] D. D. Cox and T. Dean, "Neural networks and neuroscience-inspired computer vision," *Current Biology*, vol. 24, no. 18, pp. R921 – R929, 2014.

[12] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7092–7096.

[13] B. C. Moore, L. K. Tyler, and W. Marslen-Wilson, "Introduction. the perception of speech: from sound to meaning," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 363, no. 1493, pp. 917–921, March 2008.

[14] D. Baby and H. Van hamme, "Investigating modulation spectrogram features for deep neural network-based automatic speech recognition," in *Proc. INTERSPEECH*. Dresden, Germany: ISCA, September 2015, pp. 2479–2483.

[15] A. M. C. Martinez, N. Moritz, and B. T. Meyer, "Should deep neural nets have ears? The role of auditory features in deep learning approaches," in *Proc. INTERSPEECH*. Singapore: ISCA, September 2014, pp. 2435–2439.

[16] S. Verhulst, H. M. Bharadwaj, G. Mehraei, C. A. Shera, and B. G. Shinn-Cunningham, "Functional modeling of the human auditory brainstem response to broadband stimulation," *The Journal of the Acoustical Society of America*, vol. 138, no. 3, pp. 1637–1659, Sep 2015.

[17] D. Baby and S. Verhulst, "Can biophysically inspired features improve neural network-based speech enhancement?" in *Abstract, Speech In Noise (SPIN) Workshop*, Glasgow, UK, January 2018.

[18] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. W. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proceedings of 12th International Conference on Latent Variable Analysis and Signal Separation*, Liberec, Czech Republic, August 2015, pp. 91–99.

[19] V. Hohmann, "Frequency analysis and synthesis using a Gammatone filterbank," *Acta Acustica united with Acustica*, vol. 88, no. 3, pp. 433–442, 2002.

[20] T. Irino and R. D. Patterson, "A dynamic compressive gammachirp auditory filterbank," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2222–2232, Nov 2006.

[21] R. F. Lyon, "Cascades of two-pole–two-zero asymmetric resonators are good models of peripheral auditory function," *The Journal of the Acoustical Society of America*, vol. 130, no. 6, pp. 3893–3904, 2011.

[22] A. Altoè, V. Pulkki, and S. Verhulst, "Transmission line cochlear models: Improved accuracy and efficiency," *The Journal of the Acoustical Society of America*, vol. 136, no. 4, pp. EL302–EL308, 2014.

[23] S. Verhulst, A. Altoè, and V. Vasilkov, "Computational modeling of the human auditory periphery: Auditory-nerve responses, evoked potentials and hearing loss." *Hearing Research*, vol. 360, pp. 55–75, March 2018.

[24] W. A. Dreschler, V. Hans, C. Ludvigsen, and S. Westermann, "ICRA Noises: Artificial Noise Signals with Speech-like Spectral and Temporal Properties for Hearing Instrument Assessment," *Audiology*, vol. 40, pp. 148–157, 2001.

[25] P. Søndergaard and P. Majdak, "The auditory modeling toolbox," in *The Technology of Binaural Listening*, J. Blauert, Ed. Berlin, Heidelberg: Springer, 2013, pp. 33–56.

[26] S. Bleeck, T. Ives, and R. D. Patterson, "AIM-MAT: The auditory image model in MATLAB," *Acta Acustica united with Acustica*, vol. 90, no. 4, pp. 781–787, 2004.

[27] "CARFAC Github," https://github.com/google/carfac.git, accessed: 2017-10-25.

[28] "TL 2018 Model: human cochlea + OAE + AN + ABR + EFR," https://waves.intec.ugent.be/hearing-technology, accessed: 2018-01-31.

[29] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[30] Y. Lu and P. C. Loizou, "A geometric approach to spectral subtraction," *Speech Communication*, vol. 50, no. 6, pp. 453 – 466, 2008.

[31] A. Saremi, R. Beutelmann, M. Dietz, G. Ashida, J. Kretzberg, and S. Verhulst, "A comparative study of seven human cochlear filter models," *The Journal of the Acoustical Society of America*, vol. 140, no. 3, pp. 1618–1634, 2016.