INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Causal Speech Enhancement Combining Data-driven Learning and Suppression Rule Estimation

Seyedmahdad Mirsamadi[1] *and Ivan Tashev*[2]

[1]Center for Robust Speech Systems, The University of Texas at Dallas, Richardson, TX 75080, USA
[2]Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA

mirsamadi@utdallas.edu, ivantash@microsoft.com

Abstract

The problem of single-channel speech enhancement has been traditionally addressed by using statistical signal processing algorithms that are designed to suppress time-frequency regions affected by noise. We study an alternative data-driven approach which uses deep neural networks (DNNs) to learn the transformation from noisy and reverberant speech to clean speech, with a focus on real-time applications which require low-latency causal processing. We examine several structures in which deep learning can be used within an enhancement system. These include end-to-end DNN regression from noisy to clean spectra, as well as less intervening approaches which estimate a suppression gain for each time-frequency bin instead of directly recovering the clean spectral features. We also propose a novel architecture in which the general structure of a conventional noise suppressor is preserved, but the sub-tasks are independently learned and carried out by separate networks. It is shown that DNN-based suppression gain estimation outperforms the regression approach in the causal processing mode and for noise types that are not seen during DNN training.

Index Terms: Speech enhancement, Deep neural networks, Noise suppression

1. Introduction

Single channel speech enhancement is a widely researched problem in signal processing in which the goal is to improve the perceptual quality of speech recorded in noise. While the problem of speech enhancement has been an attractive area of research in statistical signal processing for a rather long time, there has recently been new interest in replacing these algorithms with machine learning techniques which can learn the enhancement task from data.

Conventional speech enhancement algorithms rely on statistical assumptions about speech and noise signals in order to derive a *suppression rule* for each time-frequency bin, which is a real-valued gain expected to attenuate the energy of the bins that are affected by noise. These statistical noise suppression (SNS) algorithms were introduced by the pioneering studies in [1] and [2] which provide suppression rules known as Wiener rule and spectral subtraction. Another fundamental study is by Ephraim and Malah [3], where they derive an optimum Minimum Mean-Square Error (MMSE) estimator for the magnitude spectral components of the clean speech. These gain functions are optimal only when the assumed statistical model holds and the speech and noise spectral variances are known. Since these approaches rely on statistical assumptions that are often inaccurate for real data (e.g. the Gaussianity assumption for short-time spectral amplitudes or the slowly changing noise variance), most of them suffer to some extent from inadequate noise suppression (particularly for fast varying non-stationary noises), or they introduce annoying artifacts in the recovered signal [4, 5].

Followed by the success of deep neural networks in acoustic modeling for speech recognition in the last few years, there has been some interest in applying deep learning to the enhancement task where the goal is to improve the perceptual quality of the signal. Here the idea is to use a deep network architecture in a regression task to learn the complex transformation from noisy speech features to clean features. Using a DNN-based approach has the advantage that it makes no assumptions about the statistical properties of the signals, and that it can also work for fast-varying non-stationary noises (e.g. clicks, claps, etc.) because it learns frame-level transformations offline (rather than relying on past signal information to build a noise model). In [6], the authors propose to use a DNN in an end-to-end regression task where the DNN learns to transform noisy magnitude spectra to clean equivalents. They provide a comprehensive set of evaluations and show improvements compared to a conventional log-MMSE enhancement approach [7]. This work has been extended to noise-adaptive training (NAT) [8], and also further modified with variance equalization of features to alleviate the distortions in the estimated clean features [9]. The works in [10] and [11] follow very similar ideas, but use denoising auto-encoders to learn the transformation.

In this study, we employ DNN-based speech enhancement in a more realistic scenario which involves room reverberation in addition to environmental noise, and we study the case where real-time processing constraints are imposed by the application. This limits the range of context frames that can be used for the DNN input. More specifically, our goal is to design a causal DNN-based speech enhancement system to be employed in reverberant rooms, with a sufficient degree of generalization to provide reasonable performance for noise types that are not seen during training.

We examine several configurations in which DNN-based learning from simultaneous noisy/clean recordings can be incorporated into an enhancement system. This includes end-to-end regression from noisy to clean spectra, as well as architectures which estimate a suppression rule for each time-frequency bin rather than directly estimating clean features. We study two different configurations for the latter approach. One in which a single network is used to convert noisy spectra to suppression gains, and another in which the general structure of a conventional enhancement system is preserved, but the different subtasks are carried out by separate networks. We will discuss the merits of using each architecture and compare their performance on enhancing noisy and reverberant speech.

http://dx.doi.org/10.21437/Interspeech.2016-437

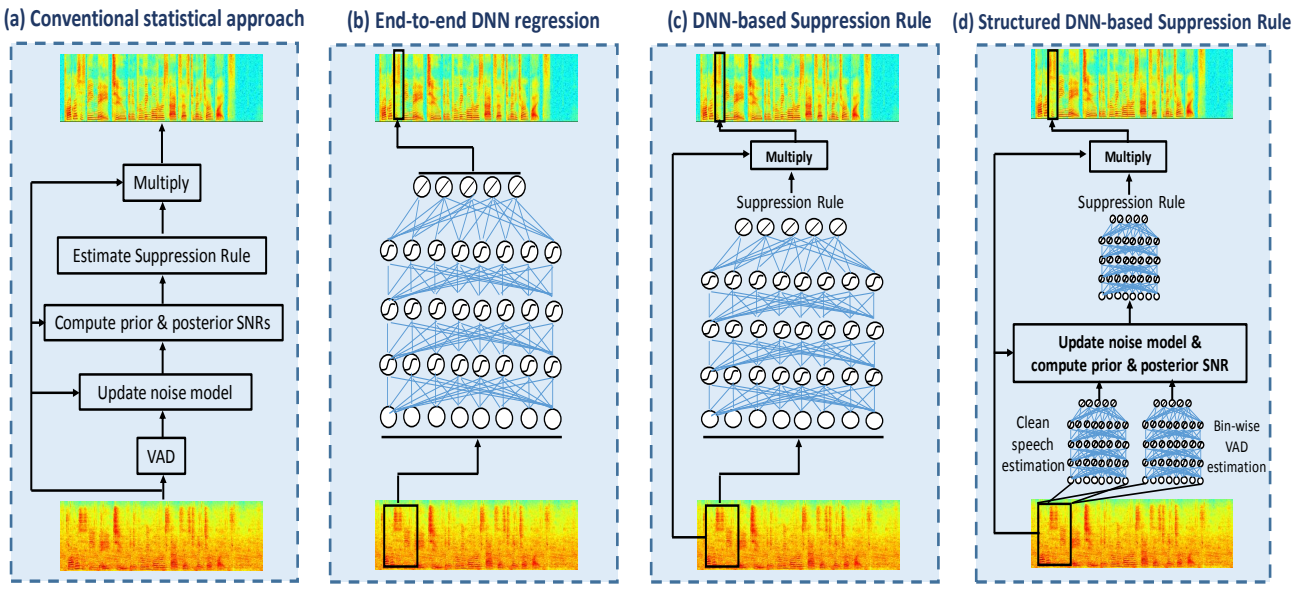

Figure 1: Different architectures for DNN-based speech enhancement: (a) Statistical noise suppression. (b) Enhancement by end-to-end DNN regression from noisy spectral features to clean features. (c) Estimating binwise suppression gain directly by a DNN. (d) Employing separate DNNs replacing the different components of conventional suppression gain estimation.

2. Conventional Speech Enhancement

Fig. 1(a) shows the major components of a typical statistical noise suppression system. A voice activity detector (VAD) estimates the probability of speech presence in each frequency bin, followed by a noise model estimation unit which updates the system's knowledge of the noise spectrum by the following recursive averaging:

$$\lambda_k(m) = \big(1-\alpha_k(m)\big)\lambda_k(m-1) + \alpha_k(m)|X_k(m)|^2, \quad (1)$$

where $\lambda_k(m)$ indicates the noise variance for time frame m and frequency bin k, $X_k(m)$ is the short-time Fourier transform (STFT) of the noisy signal, and $\alpha_k(m)$ is a recursive averaging weight which is a function of the speech presence probability for the current frame and frequency bin:

$$\alpha_k(m) = \alpha_0\big(1-p_k(m)\big)\big(1-p_{total}(m)\big). \quad (2)$$

Here, $p_k(m)$ denotes the binwise probability of speech presence (provided by the VAD module), and $p_{total}(m)$ represents the overall probability of speech presence for m'th frame, which can be obtained by averaging $p_k(m)$ across all frequency bins.

The noise spectral variance given by (1) is used to compute two quantities known as posterior SNR ($\gamma_k(m)$) and prior SNR ($\xi_k(m)$):

$$\gamma_k(m) \triangleq \frac{|X_k(m)|^2}{\lambda_k(m)} \quad (3)$$

$$\xi_k(m) \triangleq \frac{|\hat{S}_k(m)|^2}{\lambda_k(m)} \simeq \beta\xi_k(m-1)+(1-\beta)\max\{0,\gamma_k(m)-1\}, \quad (4)$$

where $\hat{S}_k(m)$ represents the clean speech spectrum, and β is a recursive averaging weight in the so-called decision-directed approach for prior SNR estimation as defined in [3]. The final suppression rule is a function of prior and posterior SNRs:

$$G_k(m) = g\big(\xi_k(m),\gamma_k(m)\big). \quad (5)$$

Different suppression rules $g(\cdot)$ have been proposed in the literature based on different statistical assumptions and optimization criteria, such as spectral subtraction [2], Wiener [12], Ephraim and Malah [3, 7] or their computationally efficient alternatives [13], as well as data-driven approaches such as [14].

3. DNN-based data-driven speech enhancement

3.1. Enhancement by DNN regression

A deep neural network can be used to learn the complex overall transformation from noisy to clean features from a dataset of synchronous clean/noisy recordings (Fig. 1(b)). The input to the network is a concatenation of magnitude STFT features from a context window of M frames which can either be symmetric (6) or causal (7):

$$\mathbf{X} = [\mathbf{x}_{m-\frac{M-1}{2}}, \cdots, \mathbf{x}_m, \cdots, \mathbf{x}_{m+\frac{M-1}{2}}], \quad (6)$$

$$\mathbf{X} = [\mathbf{x}_{m-M+1}, \cdots, \mathbf{x}_m]. \quad (7)$$

Here, $\mathbf{x}_m$ represents a vector of magnitude spectral components at different frequency bins of the observed signal for time frame m. The desired output in both symmetric and causal cases is the clean magnitude spectrum at time frame m ($\mathbf{y}_m$). Symmetric context expansion is often used in speech recognition, but for most enhancement tasks, real-time processing constraints necessitates the use of causal context expansion, in which only past frames are used as context and the goal is to recover the clean feature vector of the last frame in the context window.

3.2. DNN-based suppression rule estimation

Although the multiple layers of nonlinear transformations in an end-to-end regression can considerably remove the noise component, it also introduces distortions to the estimated clean features. Such distortions are more noticeable for unseen noise types and particularly for causal-context systems. In our experiments with causal context expansion and noise types that are *not* seen during training, although the DNN did generalize enough to leave almost no audible noise component in the output, it also severely distorted the speech component, resulting in an overall listening quality close to the original noisy speech. One approach to minimize such distortions is to estimate suppression gains at the DNN output instead of directly estimating the clean magnitude STFT (Fig. 1(c)). In other words, the desired outputs during training are set to $Y_k(m) = \frac{S_k(m)}{X_k(m)}$, where $S_k(m)$ is the magnitude STFT of the reference clean signal. The final estimated clean features are obtained as the product of the DNN's outputs and the noisy spectral features. In this case, since the relationship between enhanced and noisy spectra is forced to a simpler multiplicative relationship based on an estimated gain, the resulting artifacts are considerably reduced.

3.3. Structured suppression rule estimation using DNNs

An alternative to having a single network convert noisy spectra to suppression gains is to preserve the general architecture of a conventional speech enhancement system (VAD, noise variance estimation, suppression curve), but to learn the functionality of each of these units from data by replacing them with DNN-based substitutes (Fig. 1(d)). Such a system uses a regression DNN similar to that described in section 3.1, but instead of taking the outputs directly, they are used together with speech presence probabilities to estimate prior SNR values according to (4).

While any binwise VAD system can be used to estimate the needed speech presence probabilities, we use a second DNN which acts as VAD, providing speech presence probabilities for each time-frequency bin. This is motivated by the recent improvements reported in VAD accuracy by using a deep learning approach [15, 16]. Here, we train a simple VAD DNN in a similar fashion to the de-noising regression DNN, but the ground-truth binwise VAD labels are used as the desired outputs. Finally, instead of using fixed curves as a suppression rule, an optimum curve can be learned from data that transforms prior and posterior SNRs to the suppression gain. That is, a third DNN can be used to transform a concatenated feature vector consisting of prior and posterior SNR values at all frequencies to the corresponding suppression gain. This will have the advantage that frequency context is utilized when computing output gains (rather than having fixed curves for all frequencies). The target outputs for this network during training are set to the ratio of the clean speech component to the overall noisy speech spectral magnitude in each frequency bin (i.e. the ideal suppression gain).

While the estimation of a suppression gain curve from data has been studied before [14], our approach is different in that the prior and posterior SNR estimation step is also learned from the training data. We will show in the next section that this structured deep learning approach provides improvements particularly for the causal context and unseen noise scenarios.

4. Experiments

4.1. Datasets

A multi-condition training corpus with different noise types, signal-to-noise ratios (SNRs), and reverberant properties was created based on the TIMIT training set. We used a collection of 100 different noise signals from [17], which includes a variety of different noise types (crowd noise, traffic and car noise, etc.). We also used a set of 60 different room impulse responses (RIRs) recorded at multiple distances (from 1 to 4 meters) in a room with reverberation time (T_{60}) of approximately 300 ms. The training corpus was created as follows: speech and noise sound pressure levels (SPL) in a room were assumed to be normally distributed with means $\mu_s = 60$ dB and $\mu_n = 55$ dB, and standard deviations $\sigma_s = 8$ dB and $\sigma_n = 10$ dB. An utterance is randomly selected from the TIMIT training set, and scaled to a power level that is randomly selected according to the assumed distribution for speech power levels. Similarly, a randomly selected signal from the noise dataset is scaled to a power level chosen from the noise power distribution. The scaled speech signal is convolved with a randomly selected RIR, and the scaled noise is added to the result. This noisy signal is then synchronized with the clean speech signal to remove the delay introduced by the RIR. Such a temporal alignment of the noisy and clean reference signals is necessary so that the subsequent framing and feature extraction steps will produce feature pairs which correspond to the same section of the speech signal. The final SNRs were limited to [-5, 30] dB. This procedure is used to create a 10-hour dataset of clean/noisy pairs for training. In a similar fashion, we generate two different test datasets based on the TIMIT test set, each containing 200 utterances. The first test dataset uses the same noise signals used in the training dataset (seen noise), and the second uses a completely disjoint set of noise samples from NOISEX-92 corpus [18] (unseen noise).

4.2. System setup and configurations

As a statistical noise suppressor baseline, we use the enhancement system outlined in [19], which uses a model-based VAD [20] and spectral subtraction [2] as the suppression rule. The denoising regression DNN (Fig. 1(b)), the suppression gain estimation DNN (Fig. 1(c)) and the VAD DNN (Fig. 1(d)) all use 3 hidden layers with tanh nonlinearity and 2048 nodes per layer. The DNN which converts prior and posterior SNRs to suppression gains can be smaller, because discovering the optimum suppression rule curve is a fairly easier task compared to feature de-noising (it was experimentally verified that a single hidden layer network with 1024 nodes can effectively learn the transformation from prior and posterior SNRs to suppression gains).

A 32 ms Hann window with a skip period of 16 ms is used to segment the input signals, followed by magnitude STFT feature extraction. The input features to the DNN use a context window of 11 frames in symmetric context experiments and 7 frames in causal context experiments (these were experimentally found to provide best performance). To evaluate the performance of the different methods we use PESQ (Perceptual Evaluation of Speech Quality) [21], a computational proxy of the subjective mean opinion score (MOS) [22]. Informal listening tests have also been performed to verify the significance of the improvements.

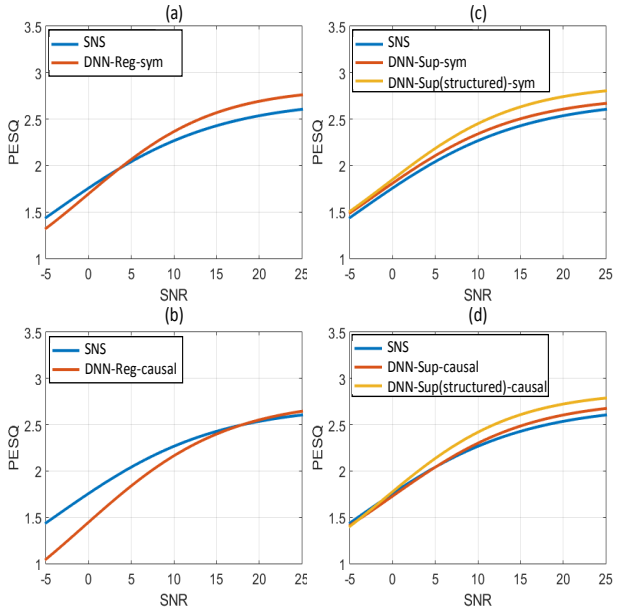

Figure 2: Output PESQ versus SNR for *unseen* noises. The blue curve representing the performance of Statistical Noise Suppressor (SNS) is the same in all 4 figures and is provided for better comparison. *Left column*: Performance of DNN regression approach in causal (a) and symmetric (b) context modes. *Right column*: Performance of DNN-based suppression gain estimation approach in causal (c) and symmetric (d) context modes.

4.3. Evaluation results

Fig. 2 shows the output PESQ versus SNR curves resulting from the different discussed methods for the unseen noise dataset. These plots have been obtained by fitting a sigmoid curve to the set of (SNR, output PESQ) pairs for the different files in the test set. In all four graphs, the performance of the statistical noise suppressor (SNS) has been shown for comparison (blue line). The two plots on the left compare the performance of end-to-end DNN regression with a statistical noise suppressor in symmetric context (Fig. 2(a)) and causal context (Fig. 2(b)) modes. The symmetric-context regression DNN can provide improvements over conventional enhancement for SNRs greater than 4 dB. However, in the causal-context mode, and also for low SNRs in the symmetric mode, end-to-end regression fails to provide satisfactory performance for unseen noises. Listening tests indicate that the reduced performance for unseen noise types (particularly in the causal-context mode) is more due to signal distortions rather than denoising ability. The two plots on the right in Fig. 2 show the performance of the DNN-based suppression gain estimation for unseen noise types (both single-DNN and structured approaches). Both approaches outperform conventional enhancement in both symmetric and causal context modes and for all SNR values. The improved PESQ values in this case is due to the considerably reduced distortions provided by using a simpler multiplicative transformation from noisy to clean features. Moreover, in our experiments for unseen noises, the structured approach almost always outperformed the single-DNN approach. We believe this is due to the explicit noise spectral variance estimation in the structured approach which complements the DNN's generalization ability to noises that are completely unseen during training.

Table 1 shows the average output PESQ over all test files in seen and unseen noise scenarios. For seen noises in symmetric context mode, end-to-end regression provides the best performance among all methods. But there is a significant performance degradation when we switch to unseen noises and causal context. In such conditions, a DNN-based suppression gain estimation outperforms both conventional enhancement and end-to-end regression. In all of the experiments, the structured DNN-based approach outperforms single-DNN suppression gain estimation particularly for unseen noise types. This is attributed to the explicit noise model estimation used in the structured approach. Also shown in Table 1 are the results of the structured approach while the third DNN (suppression curve estimation DNN) is replaced by conventional suppression curves (Wiener and spectral subtraction rules). Here, the DNN provides a marginal improvement compared to using fixed suppression curves. The third DNN is therefore just learning to mimic the function of a simple Wiener or spectral subtraction curve.

Table 1: Performance comparison of the different enhancement approaches based on resulting output PESQ.

	seen noise		unseen noise	
	sym.	causal	sym.	causal
Noisy	1.98		1.99	
SNS	2.1		2.2	
DNN (end-to-end regression)	**2.42**	2.28	2.23	2.06
DNN (suppression gain est.)	2.33	2.3	2.28	2.23
DNN (structured sup. gain est. [DNN])	2.37	2.32	2.37	**2.33**
DNN (structured sup. gain est. [Wiener])	2.4	**2.37**	**2.39**	2.31
DNN (structured sup. gain est. [SS])	2.35	2.36	2.33	2.32

5. Conclusions

We studied the use of deep neural networks for real-time low-latency speech enhancement. Several architectures were examined for DNN-based speech enhancement, including an end-to-end DNN regression from noisy to clean spectra, as well as less intervening approaches in which the DNN estimates a suppression gain similar to conventional speech enhancement. The performance of the different architectures was evaluated in both symmetric and causal context modes and for both seen and unseen noise types. While an end-to-end regression is a good choice for seen noises and symmetric context, it results in signal degradation for unseen noises and causal context, making a DNN-based suppression gain estimation a better choice in such scenarios. Also, a structured DNN-based suppression gain estimation in which the general structure of a statistical noise suppressor is preserved can outperform a single-DNN suppression gain estimation particularly in the causal-context experiments. The superior performance of the structured approach for unseen noises is due to the explicit noise variance estimation which remains in use in this architecture. However, the estimated noise variance in this case is considerably more accurate compared to conventional noise variance estimation (Equations 1 and 2), both because it uses the clean speech component from the regression DNN's output, and also because it uses more accurate speech probability estimates given by the DNN-based VAD.

6. References

[1] N. Wiener, "Extrapolation, interpolation, and smoothing of stationary time series: With engineering applications," *Principles of Electrical Engineering Series. MIT Press, Cambridge, MA*, 1949.

[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, Apr 1979.

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

[4] O. Cappé, "Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 2, pp. 345–349, Apr 1994.

[5] T. Inoue, H. Saruwatari, K. Shikano, and K. Kondo, "Theoretical analysis of musical noise in wiener filtering family via higher-order statistics," in *ICASSP 2011*, May 2011, pp. 5076–5079.

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

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443–445, Apr 1985.

[8] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "Dynamic noise aware training for speech enhancement based on deep neural networks," in *Interspeech 2014*, September 2014, pp. 2670–2674.

[9] ——, "Global variance equalization for improving deep neural network based speech enhancement," in *Signal and Information Processing (ChinaSIP), 2014 IEEE China Summit International Conference on*, July 2014, pp. 71–75.

[10] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Interspeech 2013*, September 2013, pp. 436–440.

[11] B. Xia and C. Bao, "Speech enhancement with weighted denoising autoencoder," in *Interspeech 2013*, September 2013, pp. 3444–3448.

[12] J. Lim and A. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec 1979.

[13] P. Wolfe and S. Godsill, "Simple alternatives to the ephraim and malah suppression rule for speech enhancement," in *11th IEEE Signal Processing Workshop on Statistical Signal Processing.*, 2001, pp. 496–499.

[14] J. Erkelens, J. Jensen, and R. Heusdens, "A data-driven approach to optimizing spectral speech enhancement methods for various error criteria," *Speech communication*, vol. 49, no. 7, pp. 530–541, 2007.

[15] N. Ryant, M. Liberman, and J. Yuan, "Speech activity detection on youtube using deep neural networks." in *INTERSPEECH*, 2013, pp. 728–731.

[16] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7378–7382.

[17] G. Hu, "100 nonspeech environmental sounds," [online] Available: http://web.cse.ohio-state.edu/pnl/ corpus/HuNonspeech/HuCorpus.html, 2004.

[18] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[19] I. Tashev, A. Lovitt, and A. Acero, "Unified framework for single channel speech enhancement," in *IEEE Pacific Rim Conference on Communications, Computers and Signal Processing.*, Aug 2009, pp. 883–888.

[20] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, Jan 1999.

[21] "ITU-T, recommendation p.862, perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," International Telecommunication Union-Telecommunication Standardisation Sector, Tech. Rep., 2001.

[22] "ITU-T, recommendation p.800, methods for subjective determination of transmission quality," International Telecommunication Union-Telecommunication Standardisation Sector, Tech. Rep., 2001.