JOINT NOISE AND MASK AWARE TRAINING FOR DNN-BASED SPEECH ENHANCEMENT WITH SUB-BAND FEATURES

Qing Wang[1], *Jun Du*[1], *Li-Rong Dai*[1], *Chin-Hui Lee*[2]

[1]University of Science and Technology of China, P. R. China
[2]Georgia Institute of Technology, USA

xiaosong@mail.ustc.edu.cn, {jundu,lrdai}@ustc.edu.cn, chl@ece.gatech.edu

ABSTRACT

We present a joint noise and mask aware training strategy for deep neural network (DNN) based speech enhancement with sub-band features. First, based on the analysis of the previously proposed dynamic noise aware training approach tested on the wide-band (16 KHz) speech data, the full-band dynamic noise features cannot always improve the enhancement performance due to inaccurate noise estimation. Accordingly, we improve dynamic noise estimation via enhanced post-processing, interpolation with the static noise estimation, and sub-band features. Then, the ideal ratio mask (IRM), as a relative quantity for the description of both speech and noise information, is verified to have a strong complementarity with dynamic noise estimation via joint aware training of DNN. Furthermore, a comprehensive study on different approaches to estimate noise and IRM is conducted. The experiments under unseen noises demonstrate the effectiveness of the proposed approach in both speech quality and intelligibility measures in comparison to the conventional DNN approach.

Index Terms— speech enhancement, deep neural network, dynamic noise estimation, ideal ratio mask, sub-band features

1. INTRODUCTION

Speech enhancement techniques have became extremely important in real-world applications, such as automatic speech recognition (ASR), mobile communications, and hearing aids [1]. The speech enhancement performance in real acoustic environments is not always satisfactory due to the complexity of noise corruption on speech. The conventional speech signal processing methods, e.g., spectral subtraction [2], Wiener filtering [3], minimum mean squared error (MMSE) estimation [4, 5] and optimally-modified log-spectral amplitude (OM-LSA) speech estimator [6] have been proposed during the past several decades. Model assumptions for the interactions between speech and noise are made in these methods, which often lead to the failure of tracking non-stationary noises for real-world scenarios in unexpected acoustic conditions and musical noise artifacts [7].

Recently, with the fast development of deep learning techniques [8, 9], the deep architecture was adopted to model the complicated relationship between noisy speech and clean speech in speech enhancement area [10, 11, 12, 13]. Previously we proposed a deep neural network (DNN) based speech enhancement framework to map noisy log-power spectra (LPS) features to clean LPS features [14, 15]. And a large number of different noise types could be included in the training set to alleviate the mismatch problem between training and testing. In [16], many different kinds of noise types were also used to train DNNs to predict the ideal binary mask (IBM), and the robustness to unseen noise types was demonstrated. Therefore, one advantage of DNN-based speech enhancement method is that the relationship between noisy speech and clean speech could be well learned from the large-scale multi-condition data.

Furthermore, it was verified [15, 17] that the static noise information estimated by the first several noise frames of the utterance, namely the static noise aware training (SNAT), can make a better prediction of the clean speech and suppression of the additive noises. To handle the non-stationary or burst noises, the dynamic noise aware training (DNAT) approach was proposed [18]. However, due to the inaccurate estimation of dynamic noise information, the performance is not always satisfactory. Accordingly, this study first improves dynamic noise estimation via enhanced post-processing, sub-band features, and interpolation with the static noise estimation. Then, the ideal ratio mask (IRM), as a relative quantity for the description of both speech and noise information, is verified to have a strong complementarity with dynamic noise estimation via joint aware training of DNN. Finally, a comprehensive study on different approaches to estimate noise and IRM is conducted. The experiments under unseen noises demonstrate the effectiveness of the proposed approach in both speech quality and intelligibility measures in comparison to the conventional DNN approach.

In Section 2, the DNN architecture is introduced. In Section 3, improved dynamic noise estimation is presented. In Section 4, joint noise and mask aware training is described. In Section 5 and 6, we give experiments and conclusions.

2. THE DNN ARCHITECTURE

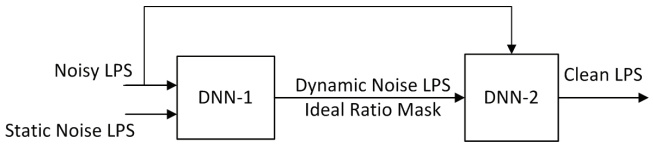

Fig. 1. The proposed DNN-based framework.

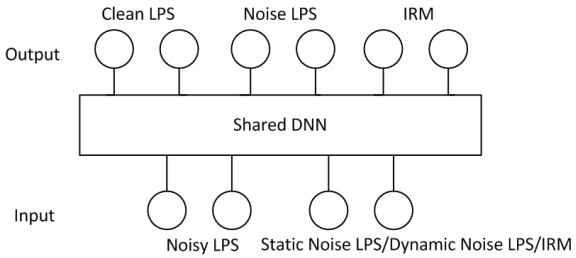

Fig. 2. The DNN architecture.

A block diagram of the proposed speech enhancement framework is illustrated in Fig. 1. Two regression DNNs (denoted as DNN-1 and DNN-2), similar to [15], should be built. First, DNN-1 aims to provide dynamic noise and IRM estimation. With both noisy LPS features and static noise LPS features as the input, DNN-1 refers to SNAT system [18]. Then DNN-2 can perform joint noise and mask aware training to make a better prediction of the clean LPS features. The general architecture using multiple outputs for both DNN-1 and DNN-2 is illustrated in Fig. 2. The MMSE criterion is adopted to optimize the DNN parameters as follows:

$$E = \sum_{t=1}^{T} \left(\|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_2^2 + \alpha \|\hat{\mathbf{n}}_t - \mathbf{n}_t\|_2^2 + \beta \|\hat{\mathbf{m}}_t - \mathbf{m}_t\|_2^2 \right) \quad (1)$$

where $\hat{\mathbf{x}}_t$ and $\mathbf{x}_t$ are the t^{th} D_1-dimensional vectors of estimated and clean reference LPS features, respectively, with T representing the mini-batch size. $\mathbf{n}_t$ is the t^{th} D_2-dimensional reference noise LPS sub-band features while $\mathbf{m}_t$ is the t^{th} D_2-dimensional IRM sub-band features. α and β are the weighting coefficients. The linear activation function is used for clean and noise outputs while the sigmoid activation function is adopted for the IRM output. As shown in Table 1, several DNN systems using noise or IRM aware training will be compared with different settings of DNN-1 and DNN-2 architectures. **SNAT** and **DNAT** are the static and dynamic noise aware training systems in [18]. **IDNAT** is the improved DNAT system described in Section 3. Both **DNAT** and **IDNAT** use the single output architecture ($\alpha = \beta = 0$) for DNN-1. **MAT** denotes the system with IRM aware training, where the dual output architecture ($\alpha = 0, \beta \neq 0$) is adopted for DNN-1 to provide the IRM estimation. **JAT** represents the system with joint noise and mask aware training, where the triple output architecture ($\alpha \neq 0, \beta \neq 0$) is designed for DNN-1. For all systems, the single output architecture is always used for DNN-2. The other details of proposed speech enhancement system, including DNN training/decoding, feature extraction and waveform reconstruction, can refer to [14, 15, 19].

System	DNN-1		DNN-2	
SNAT	$\alpha = 0$	$\beta = 0$	-	-
DNAT/IDNAT	$\alpha = 0$	$\beta = 0$	$\alpha = 0$	$\beta = 0$
MAT	$\alpha = 0$	$\beta = 0.05$	$\alpha = 0$	$\beta = 0$
JAT	$\alpha = 0.05$	$\beta = 0.05$	$\alpha = 0$	$\beta = 0$

Table 1. The setting of DNN-1/DNN-2 for different systems.

3. IMPROVED DYNAMIC NOISE AWARE TRAINING

In [18], both SNAT and DNAT have been investigated. And the experiments on the narrow-band (8 kHz) speech data showed that DNAT is more effective than SNAT. However, based on the analysis on the wide-band (16 kHz) speech data, the full-band dynamic noise LPS features cannot always improve the enhancement performance due to inaccurate noise estimation, which might be explained as that the relationship between the noisy speech and clean speech in higher dimensional feature space is much more challenging for DNN to handle. To address this problem, three strategies are proposed to improve the dynamic noise estimation.

3.1. Enhanced post-processing for noise estimation

The frame-level dynamic noise estimation in [18] was implemented via the post-processing of estimated clean speech from DNN-1 output. First, a ratio γ between the estimated clean speech and input noisy speech in the power spectral domain is defined as:

$$\gamma(d) = \exp\left(\hat{x}_t(d) - y_t(d)\right) \quad (2)$$

where $\hat{x}_t(d)$ is the d^{th} element of estimated clean speech LPS feature vector $\hat{\mathbf{x}}_t$ and $y_t(d)$ is the corresponding version of input noisy speech. Then an IBM can be estimated by a global threshold λ. However, this estimation is not robust to the cases that the absolute energy of the time-frequency (T-F) bin is quite high or low. Accordingly, we design a new IBM estimation method:

$$\hat{\text{IBM}}_t(d) = \begin{cases} 1 & \gamma(d) > \lambda \text{ and } \hat{x}_t(d) > E_t^l \\ 0 & \gamma(d) > \lambda \text{ and } \hat{x}_t(d) \leq E_t^l \\ 1 & \gamma(d) \leq \lambda \text{ and } \hat{x}_t(d) > E_t^h \\ 0 & \gamma(d) \leq \lambda \text{ and } \hat{x}_t(d) \leq E_t^h \end{cases} \quad (3)$$

where E_t^h and E_t^l are high and low thresholds of LPS features at the t^{th} frame which are calculated as:

$$\begin{aligned} E_t^h &= E_t + E_h \\ E_t^l &= E_t + E_l \end{aligned} \quad (4)$$

where E_t is an adaptive threshold averaged in a context window size of 11-frame estimated clean LPS features. E_h and E_l are the fixed high and low thresholds. The idea of using double thresholds is inspired by the work in voice activity detection [20]. Furthermore, the IBM from Eq. (3) can be smoothed in each T-F bin with a context window size of 5 frames. Finally, the noise estimation based on IBM is the same as that in [18].

3.2. Interpolation of static and dynamic noise estimation

Another strategy to alleviate the problem of inaccurate dynamic noise estimation is to perform a linear interpolation between static and dynamic noise estimation:

$$\hat{\mathbf{n}}_t^{\text{new}} = \frac{1}{2}\left(\hat{\mathbf{n}}^{\text{S}} + \hat{\mathbf{n}}_t^{\text{D}}\right) \tag{5}$$

which is motivated by the complementarity between them, namely the static noise estimation is a stable representation of noise statistics while the dynamic noise estimation corresponds to the details of noise statistics in each frame.

3.3. Sub-band features

Inspired by the success of DNAT on the 8 kHz speech data, a straightforward way is to reduce the high dimension of the estimated noise LPS feature vector. Thus, we design the sub-band features by mapping the linear frequency bins of D_1-dimensional ($D_1 = 257$) full-band LPS features to frequency bins of D_2 ($D_2 = 64$) gammatone filter banks which can simulate the frequency selectivity of human ears [21], as illustrated in Fig. 3. In each sub-band, the mapped feature can be computed as:

$$\hat{n}_t^{\text{sub}}(i) = \frac{\sum_{d_i \le d < d_{i+1}} \hat{n}_t^{\text{full}}(d)}{d_{i+1} - d_i}, i = 1, 2, \ldots, D_2 \tag{6}$$

where d_i is the starting index of the i^{th} sub-band. The sub-band noise features not only improve the enhancement performance but also reduce the model size and the computational complexity of DNN.

4. JOINT NOISE AND MASK AWARE TRAINING

IRM [22, 23] is a measure to estimate the speech presence in a local T-F unit, which is extended from the IBM widely used in computational auditory scene analysis (CASA). As a soft mask, IRM can achieve better speech separation performance [24], which can be implemented as:

$$m_t(d) = \frac{\exp\left(x_t(d)\right)}{\exp\left(x_t(d)\right) + \exp\left(n_t(d)\right)} \tag{7}$$

where the $\exp(\cdot)$ operation transforms the LPS features back to the linear frequency domain. As the mask is highly related with the auditory attention mechanism, the mask aware

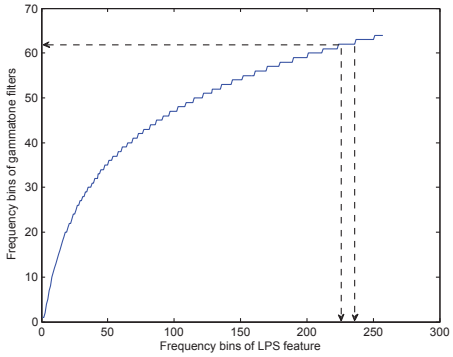

Fig. 3. Illustration of the mapping between full-band (257-dimension) and sub-band (64-dimension) LPS features.

training can be treated as the implicit attention-based DNN training where IRM is an indicator of speech presence or absence. However, according to the preliminary experiments, MAT using only IRM information can not significantly improve the performance. So we design a joint noise and mask aware training approach by concatenating both the dynamic noise estimation and IRM with the input noisy speech features:

$$\mathbf{z}_t = \left[\mathbf{y}_{t-\tau}^{t+\tau}, \hat{\mathbf{n}}_t, \hat{\mathbf{m}}_t\right] \tag{8}$$

where $\mathbf{z}_t$ is the input vector of DNN-2. $\mathbf{y}_{t-\tau}^{t+\tau}$ denotes the input noisy speech LPS feature vector with $2\tau+1$ frame expansion. $\hat{\mathbf{m}}_t$ is one output of the DNN-1 and $\hat{\mathbf{n}}_t$ is calculated according to Section 3. Please note that $\hat{\mathbf{m}}_t$ also uses the sub-band features with $D_2 = 64$. We believe that IRM as a relative quantity for the description of both speech and noise information could be complementary with the dynamic noise estimation to better predict the clean speech.

5. EXPERIMENTAL RESULTS AND ANALYSIS

In this work, we extended sample rate of waveforms from 8 kHz [18] to 16 kHz. 115 noise types including 100 noise types in [25] and some other musical noises were adopted to improve the generalization capacity of DNN. All 4620 utterances from the training set of the TIMIT database [26] were corrupted with the abovementioned 115 noise types at six levels of SNR, i.e., 20dB, 15dB, 10dB, 5dB, 0dB, and -5dB, to build the multi-condition training set. We randomly selected a 10-hour training set with 11550 utterance pairs. The 192 utterances from core test set of TIMIT database were used to construct the test set. Three unseen noise types, namely Buccaneer1, Destroyer engine and Leopard from the NOISEX-92 corpus [27], were adopted for testing.

The frame length was set to 512 samples (32 msec) with a frame shift of 256 samples. With short-time Fourier analysis, 257-dimensional LPS features [19] were obtained to train DNNs. Mean and variance normalization were applied to the

input and target feature vectors of the DNN. All DNN configurations were fixed at 3 hidden layers, 2048 units for each hidden layer and 7-frame input. For SNAT system, the first 6 frames of each utterance were used for noise estimation. For dynamic noise estimation, the λ was set to 0.1. E_h and E_l were set to 4 and -1 respectively. Perceptual evaluation of speech quality (PESQ) [28] and short-time objective intelligibility (STOI) [29] were used to assess the quality and intelligibility of the enhanced speech.

5.1. Evaluation on SNAT and DNAT

Table 2 lists the performance comparison of several systems mentioned in [18] on the 16 kHz speech data. The DNN baseline system with only noisy speech LPS features as the input significantly improved the PESQ and STOI over the original noisy speech. And the SNAT system consistently outperformed DNN baseline system. One exception was that DNAT underperformed SNAT which was not consistent with the observation in [18], which was explained as that the relationship between the noisy speech and clean speech in higher dimensional feature space was much more challenging for DNN to learn. This led to the inaccurate noise estimation in frame-level.

	Noisy		DNN baseline		SNAT		DNAT	
SNR(dB)	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI
20	3.070	0.962	3.441	0.948	3.560	0.955	3.546	0.948
15	2.727	0.924	3.207	0.928	3.310	0.935	3.316	0.929
10	2.397	0.866	2.942	0.895	3.022	0.904	3.038	0.899
5	2.073	0.788	2.630	0.843	2.705	0.857	2.694	0.850
0	1.755	0.698	2.275	0.766	2.344	0.789	2.296	0.776
-5	1.470	0.608	1.894	0.667	1.952	0.699	1.862	0.672
Ave	2.249	0.808	2.732	0.841	2.815	0.856	2.792	0.846

Table 2. PESQ and STOI comparison of different systems on the test set averaged on three unseen noises.

5.2. Evaluation on IDNAT

Based on the analysis of DNAT results, Table 3 progressively shows the performance improvements of three strategies of IDNAT. "+EnhPP" improved DNAT via the enhanced post-processing. "+Interpolation" further adopted the interpolation of static and dynamic noise estimation. "+Subband" used all three strategies, namely the IDNAT system. Obviously, the enhanced post-processing was mainly effective for the low SNR cases. Both the interpolation and sub-band features consistently yielded performance gains for all SNRs and measures (only one exception for STOI under -5dB). Overall, the IDNAT system achieved an average PESQ gain of 0.1 and an average STOI gain of 0.01 over the DNAT system.

5.3. Evaluation on MAT and JAT

Finally, Table 4 gives performance comparison of MAT and JAT systems. JAT-1 system used the dynamic noise estimation from the one output of DNN-1 while JAT-2 system adopted the method in Section 3 to estimate the dynamic noise. The MAT system using IRM information achieved comparable performance with SNAT and IDNAT system which demonstrated the effectiveness of the IRM as an auditory attention mechanism to guide the DNN training. JAT-2 obtained better PESQ performance than JAT-1, indicating the improved dynamic noise estimation was more stable than the learned noise information. In comparison to the best SNAT results in Table 2, the JAT-2 system significantly improved both speech quality and intelligibility with an average PESQ gain of 0.137 and an average STOI gain of 0.016.

	+EnhPP		+Interpolation		+Subband	
SNR(dB)	PESQ	STOI	PESQ	STOI	PESQ	STOI
20	3.529	0.945	3.556	0.948	3.604	0.955
15	3.311	0.928	3.334	0.931	3.376	0.938
10	3.049	0.899	3.070	0.902	3.106	0.908
5	2.736	0.853	2.753	0.855	2.789	0.861
0	2.380	0.784	2.387	0.788	2.432	0.790
-5	1.994	0.687	1.997	0.693	2.031	0.691
Ave	2.833	0.849	2.849	0.853	2.890	0.857

Table 3. PESQ and STOI comparison of three strategies for IDNAT system on the test set averaged on three unseen noises.

	MAT		JAT-1		JAT-2	
SNR(dB)	PESQ	STOI	PESQ	STOI	PESQ	STOI
20	3.557	0.957	3.638	0.965	3.654	0.965
15	3.318	0.938	3.404	0.947	3.422	0.947
10	3.044	0.908	3.129	0.918	3.151	0.918
5	2.731	0.861	2.819	0.873	2.848	0.874
0	2.387	0.794	2.475	0.808	2.507	0.809
-5	2.010	0.704	2.108	0.717	2.129	0.718
Ave	2.841	0.860	2.929	0.871	2.952	0.872

Table 4. PESQ and STOI comparison of MAT and JAT systems on the test set averaged on three unseen noises.

6. CONCLUSION

We propose a joint noise and mask aware training strategy for DNN-based speech enhancement with sub-band features. The inaccurate noise estimation problem of DNAT is alleviated via the IDNAT. And JAT can significantly outperform IDNAT and MAT which indicates the strong complementarity between dynamic noise estimation and IRM information.

7. ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Grants No. 61671422, National Key Technology Support Program under Grants No. 2014BAK15B05, the "Strategic Priority Research Program" of the Chinese Academy of Sciences under Grant No. XDB02070006.

8. REFERENCES

[1] J. Benesty, S. Makino, and J. D. Chen, *Speech Enhancement*, Springer, 2005.

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

[3] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Transactions on Acoustic, Speech, Signal Processing*, vol. 26, no. 3, pp. 197–210, 1978.

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

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

[6] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[7] A. Hussain, M. Chetouani, S. Squartini, A. Bastari, and F. Piazza, *Nonlinear Speech Enhancement: An Overview*, Springer, 2007.

[8] Y. Bengio, "Learning deep architectures for ai," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[9] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[10] A. L. Maas, T. M. ONeil, A. Y. Hannun, and A. Y. Ng, "Recurrent neural network feature enhancement: The 2nd chime challenge," in *Proceedings The 2nd CHiME Workshop on Machine Listening in Multisource Environments held in conjunction with ICASSP*, 2013, pp. 79–80.

[11] A. L. Maas, Q. V. Le, T. M. ONeil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust asr," in *Proc. Interspeech*, 2012.

[12] B. Xia and C. Bao, "Speech enhancement with weighted denoising auto-encoder," in *Proc. Interspeech*, 2013, pp. 3444–3448.

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

[14] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65, 2014.

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

[16] Y. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

[17] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. of ICASSP*, 2013, pp. 7398–7402.

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

[19] J. Du and Q. Huo, "A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions," in *Proc. Interspeech*, 2008, pp. 569–572.

[20] P. Renevey and A. Drygajlo, "Entropy based voice activity detection in very noisy conditions," in *Proc. EUROSPEECH*, 2001, pp. 1887–1890.

[21] D. L. Wang and G. J. Broun, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, Wiley/IEEE Press, 2006.

[22] D.L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*, pp. 181–197. Springer, 2005.

[23] S. Srinivasan, N. Roman, and D. L. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication*, vol. 48, no. 11, pp. 1486–1501, 2006.

[24] C. Hummersone, T. Stokes, and T. Brookes, "On the ideal ratio mask as the goal of computational auditory scene analysis," in *Blind Source Separation*, pp. 349–368. Springer, 2014.

[25] G. Hu, "100 nonspeech environmental sounds," 2014.

[26] J. S Garofolo et al., "Getting started with the darpa timit cd-rom: An acoustic phonetic continuous speech database," *National Institute of Standards and Technology (NIST), Gaithersburgh, MD*, vol. 107, 1988.

[27] 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.

[28] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. of ICASSP*, 2001, vol. 2, pp. 749–752.

[29] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.