ENHANCED TIME-FREQUENCY MASKING BY USING NEURAL NETWORKS FOR MONAURAL SOURCE SEPARATION IN REVERBERANT ROOM ENVIRONMENTS

Yang Sun[1], *Wenwu Wang*[2], *Jonathon A. Chambers*[1], *Syed Mohsen Naqvi*[1]

[1]Intelligent Sensing and Communications Research Group, Newcastle University, UK
[2]Centre for Vision, Speech and Signal Processing, University of Surrey, UK

ABSTRACT

Deep neural networks (DNNs) have been used for dereverberation and denosing in the monaural source separation problem. However, the performance of current state-of-the-art methods is limited, particularly when applied in highly reverberant room environments. In this paper, we propose an enhanced time-frequency (T-F) mask to improve the separation performance. The ideal enhanced mask (IEM) consists of the dereverberation mask (DM) and the ideal ratio mask (IRM). The DM is specifically applied to eliminate the reverberations in the speech mixture and the IRM helps in denoising. The IEEE and the TIMIT corpora with real room impulse responses (RIRs) and noise from the NOISEX dataset are used to generate speech mixtures for evaluations. The proposed method outperforms the state-of-the-art methods specifically in highly reverberant and noisy room environments.

Index Terms— source separation, reverberant room environments, dereverberation, time-frequency mask

1. INTRODUCTION

Speech separation aims to extract the target speech signal from the mixture which contains the background interferences [1, 2]. In real room environments, the target speech signal and background interferences have reflections which affect the perceptual quality and intelligibility of the target speech signal. Meanwhile, in many applications such as automatic speech recognition (ASR), assisted living systems and hearing aids, if the undesired signals and their reflections are removed from the mixture, the capabilities in these applications will be further improved [3–5].

Many approaches have been developed to solve the source separation problem in monaural and binaural cases [6–8]. In recent studies, the masking-based DNN method is applied to predict a T-F mask to separate the target speech from the mixture containing noise and reverberations [4,8,9]. When the T-F mask is applied to the mixture, the speech-dominant parts are preserved and the noise-dominant parts are suppressed, hence, the target speech is separated. Generally, the training targets of the masking-based DNNs are classified as ideal binary mask (IBM), IRM and complex IRM (cIRM).

In IBM, each T-F unit is assigned as 1 or 0 according to the criterion for the active source [10]. In IRM, each T-F unit is a ratio between the energies of the target speech signal and the mixture [11]. However, the limitation of the IRM is that the phase information of the clean speech signal is not used in speech reconstruction. To overcome this drawback, the cIRM is proposed, where the phase information of the speech mixture is considered [12]. The cIRM is a complex T-F mask which is obtained by using the real and imaginary components of the short-time Fourier transform (STFT). According to [11], by using the IRM to separate the speech mixture, the separation performance is always better than using the IBM. Hence, the IRM and the cIRM are used for the performance comparison with our proposed method. However, in real-world environments, the separation performance of the above mentioned methods is limited, not always yielding robust performance in various environments and noise is still challenging.

In this paper, we propose a new dereverberation and separation method. Firstly, a DNN is trained to generate the DM, which is applied to eliminate the reflections. Then, the DM is integrated with the IRM for final separation of the speech mixture. The paper is organized as follows. Section 2 describes the IEM and the framework of the proposed method. The experimental results and analysis are shown in Section 3. The conclusions and future work are given in Section 4.

2. PROPOSED METHOD

Assume that $s(m)$, $n(m)$ and $y(m)$ are the target speech signal, the noise and the acquired mixture at discrete time m, respectively. The $h_s(m)$ and $h_n(m)$ are the RIRs for reverberant speech and noise, respectively. The convolutive mixture is expressed as:

$$y(m) = s(m) * h_s(m) + n(m) * h_n(m) \tag{1}$$

where '$*$' indicates the convolution operator.

By using the STFT, the mixture is written as:

$$Y(t,f) = S(t,f)H_s(t,f) + N(t,f)H_n(t,f) \tag{2}$$

where $S(t,f)$, $N(t,f)$ and $Y(t,f)$ are the spectra of speech, noise and mixture, respectively. The $H_s(t,f)$ and $H_n(t,f)$

ISBN 978-90-827970-1-5

are the RIRs for speech and noise at time frame t and frequency f.

The target of the dereverberation is to remove the reflections in the reverberant mixture and obtain only anechoic mixture. Because the DNN can be utilized to model the relationship between the input of the DNN and the training target, we proposed the DNN-based method to achieve the dereverberation.

According to (2), we rewrite the reverberant mixture as:

$$Y(t,f) = \big(S(t,f)+N(t,f)\big)\left(\frac{H_s(t,f)}{1+\frac{N(t,f)}{S(t,f)}} + \frac{H_n(t,f)}{1+\frac{S(t,f)}{N(t,f)}}\right) \tag{3}$$

Therefore, by using the $Y(t,f)$ and $\big(S(t,f)+N(t,f)\big)$, the relationship between the reverberant and anechoic mixtures is obtained.

Hence, in our proposed method, we defined the dereverberation mask (DM) as:

$$DM(t,f) = \left(\frac{H_s(t,f)}{1+\frac{N(t,f)}{S(t,f)}} + \frac{H_n(t,f)}{1+\frac{S(t,f)}{N(t,f)}}\right)^{-1} \tag{4}$$

In the training stage, the spectra of speech $S(t,f)$, noise $N(t,f)$ and reverberant mixture $Y(t,f)$ are available, therefore, the DM can be learned as:

$$DM(t,f) = \big(S(t,f)+N(t,f)\big)Y(t,f)^{-1} \tag{5}$$

By using (3) to (5), we can obtain the anechoic mixture which is separated with the ideal ratio mask (IRM). As in [11], we define the IRM:

$$IRM(t,f) = \left(\frac{|S(t,f)|^2}{|S(t,f)|^2+|N(t,f)|^2}\right)^{\beta} \tag{6}$$

where β is a tunable parameter to scale the mask, $|S(t,f)|$ and $|N(t,f)|$ denote the speech and noise magnitude spectrum, respectively.

Therefore, the ideal enhanced mask (IEM) can be generated by integrating the DM and the IRM as:

$$IEM(t,f) = DM(t,f)IRM(t,f) \tag{7}$$

According to (4) and (5), we see that the DM is a dereverberation operation. Thus, we have

$$S(t,f)+N(t,f) = Y(t,f)DM(t,f) \tag{8}$$

The dereverberation and separation are jointly achieved with the IEM and the separated speech signal is expressed as:

$$S(t,f) = Y(t,f)IEM(t,f) \tag{9}$$

Hence, the DM is used to achieve the dereverberation and the IRM is employed to separate the target speech. The values in the DM have a large range, therefore, the compression and recovery processes are essential. In the training stage, the compressed IEM, $M_c(t,f)$ is written as:

$$M_c(t,f) = V\frac{1-e^{-C\cdot IEM(t,f)}}{1+e^{-C\cdot IEM(t,f)}} \tag{10}$$

where C is the steepness constraint and the value of $M_c(t,f)$ is limited in the range $[-V,V]$. In the proposed method, the training objective of the DNN is the compressed IEM, which is calculated based on feature combinations [11]. After the validation tests, the values of C and V are chosen as 1 and 10, respectively.

In the testing stage, the output of the trained DNN is recovered and the final predicted T-F mask is expressed as:

$$\hat{M}(t,f) = -\frac{1}{C}\log\Big(\frac{V-O(t,f)}{V+O(t,f)}\Big) \tag{11}$$

where the $\hat{M}(t,f)$ is the predicted IEM, and $O(t,f)$ is the output of the trained DNN.

Therefore, the predicted target signal $\hat{S}$ is obtained by using the final predicted mask:

$$\hat{S}(t,f) = Y(t,f)\hat{M}(t,f) \tag{12}$$

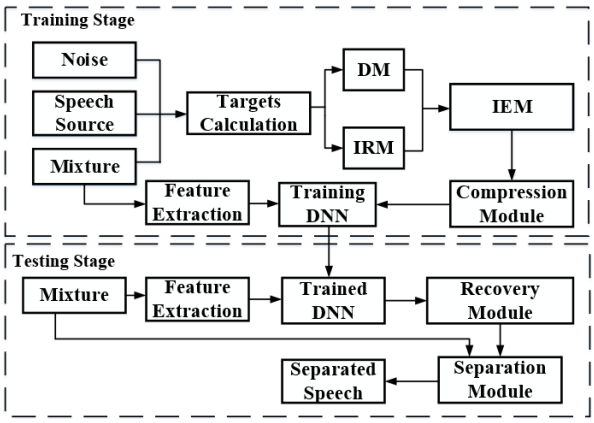

Fig. 1: The block diagram of the proposed method. The trained DNN is given by the training stage and the output of the testing stage is the separated speech signal.

Figure 1 is the flow diagram of our proposed method, where (10) and (11) are achieved in the compression module and the recovery module, respectively. The target speech signal is separated from the convolutive mixture with the predicted IEM in the separation module.

3. EVALUATIONS AND RESULTS

The speech sources are selected from the IEEE [13] and the TIMIT corpora [14]. The noise signals are selected from the NOISEX database [15] and the real RIRs [16] are used. The speech utterances are mixed with four types of noise and room

ISBN 978-90-827970-1-5 © EURASIP 2018

Table 1: Separation performance comparison in terms of SNR_{fw} (dB) with different training targets, SNR levels and RT60s. The noise in the experiments is **factory** noise. Each result is the average value of 120 experiments. **BOLD** indicates the best result.

RT60s	Room A (0.32 s)			Room B (0.47 s)			Room C (0.68 s)			Room D (0.89 s)		
SNR Levels (dB)	-3	0	3	-3	0	3	-3	0	3	-3	0	3
Mixture	1.82	2.36	3.02	1.36	1.79	2.33	1.41	1.92	2.01	0.91	1.22	1.59
IRM [11]	3.78	4.38	4.95	4.25	5.06	5.66	4.78	5.57	6.27	3.58	4.12	4.53
cIRM [12]	4.12	4.72	5.12	4.53	5.25	5.79	4.81	5.77	6.35	3.98	4.48	5.01
Proposed	**4.70**	**5.39**	**6.06**	**4.61**	**5.42**	**6.11**	**5.16**	**5.97**	**6.62**	**4.68**	**5.27**	**5.91**

Table 2: Separation performance comparison in terms of SNR_{fw} (dB) with different training targets, SNR levels and RT60s. The noise in the experiments is **babble** noise. Each result is the average value of 120 experiments. **BOLD** indicates the best result.

RT60s	Room A (0.32 s)			Room B (0.47 s)			Room C (0.68 s)			Room D (0.89 s)		
SNR Levels (dB)	-3	0	3	-3	0	3	-3	0	3	-3	0	3
Mixture	1.90	2.50	3.18	1.44	1.94	2.50	0.69	1.18	1.67	0.69	1.08	1.47
IRM [11]	4.28	4.67	5.08	4.97	5.34	5.96	5.38	5.99	6.65	3.80	4.31	4.75
cIRM [12]	4.88	5.05	5.43	4.97	5.43	6.40	4.89	5.66	6.75	4.28	4.51	4.79
Proposed	**5.33**	**5.64**	**6.01**	**5.10**	**5.56**	**6.61**	**5.61**	**6.29**	**7.21**	**5.00**	**5.52**	**5.99**

Table 3: Separation performance comparison in terms of SNR_{fw} (dB) with different training targets, SNR levels and RT60s. The noise in the experiments is **cafe** noise. Each result is the average value of 120 experiments. **BOLD** indicates the best result.

RT60s	Room A (0.32 s)			Room B (0.47 s)			Room C (0.68 s)			Room D (0.89 s)		
SNR Levels (dB)	-3	0	3	-3	0	3	-3	0	3	-3	0	3
Mixture	3.03	3.56	4.11	2.28	2.60	2.91	2.49	2.99	3.52	2.03	2.20	2.39
IRM [11]	4.13	4.65	5.23	4.59	5.35	5.98	5.27	6.13	6.77	3.64	4.03	4.43
cIRM [12]	4.62	5.01	5.43	4.96	5.72	6.01	5.26	6.00	6.07	4.18	4.53	5.11
Proposed	**5.14**	**5.95**	**6.43**	**5.07**	**5.98**	**6.14**	**5.53**	**6.58**	**7.34**	**4.84**	**5.53**	**6.05**

Table 4: Separation performance comparison in terms of SNR_{fw} (dB) with different training targets, SNR levels and RT60s. The noise in the experiments is **SSN** noise. Each result is the average value of 120 experiments. **BOLD** indicates the best result.

RT60s	Room A (0.32 s)			Room B (0.47 s)			Room C (0.68 s)			Room D (0.89 s)		
SNR Levels (dB)	-3	0	3	-3	0	3	-3	0	3	-3	0	3
Mixture	2.88	3.27	3.71	2.39	2.75	3.56	2.56	2.88	3.33	2.17	2.41	2.67
IRM [11]	5.50	5.61	6.04	5.00	5.30	5.46	6.59	7.18	7.60	4.89	5.24	5.31
cIRM [12]	5.52	5.54	6.17	5.21	5.53	5.69	5.69	6.50	6.97	5.07	5.44	5.67
Proposed	**6.17**	**6.49**	**6.80**	**6.02**	**6.39**	**6.55**	**6.87**	**7.96**	**8.25**	**5.98**	**6.31**	**6.71**

environments, which have different RT60s. In these noise signals, a speech-shaped noise (SSN) is generated as the stationary noise [17] and all others are the non-stationary noise, namely factory, babble and cafe. In the experiments, we randomly select 780, 100 and 120 utterances from the IEEE and the TIMIT corpora. These clean utterances are used to mix with noise at the different signal-to-noise ratio (SNR) levels and RIRs to generate training, development and testing datasets. Besides, for each room, five different RIRs are used to train the DNN and in the testing data, all of the mixtures with these five RIRs are evaluated. The azimuth between two signal sources are selected from 0° to 60° with 15° increment. The numbers of mixtures in training, development and testing data for each room are 46800, 6000 and 7200, respectively. Table 5 illustrates the parameters in the real RIRs [16].

The DNNs are trained by using the AdaGrad algorithm with a momentum term for 100 epochs. The learning rate is linearly decreased from 1 to 0.01, while the momentum is fixed as 0.9 in the first ten epochs and changed as 0.5 till the end. We compare the proposed method with two state-of-the-art T-F masks: the IRM [11] and the cIRM [8]. The evaluation measures are the frequency-weighted segmental SNR (SNR_{fw}) [18] and the source to distortion ratio (SDR) [19].

Table 5: The parameters for real RIRs in different rooms [16]

Room	Size	Dimension (m^3)	$RT60$ (s)
A	Medium	$5.7 \times 6.6 \times 2.3$	0.32
B	Small	$4.7 \times 4.7 \times 2.7$	0.47
C	Large	$23.5 \times 18.8 \times 4.6$	0.68
D	Medium	$8.0 \times 8.7 \times 4.3$	0.89

In our experiments, the DNN has three hidden layers and each hidden layer has 1024 units. The activation function for each hidden unit is selected as the rectified linear unit (ReLU) to avoid the gradient vanishing problem and the output layer has linear units [8]. The context window is employed to utilize the temporal information between neighbouring frames and the window length is three [11].

ISBN 978-90-827970-1-5

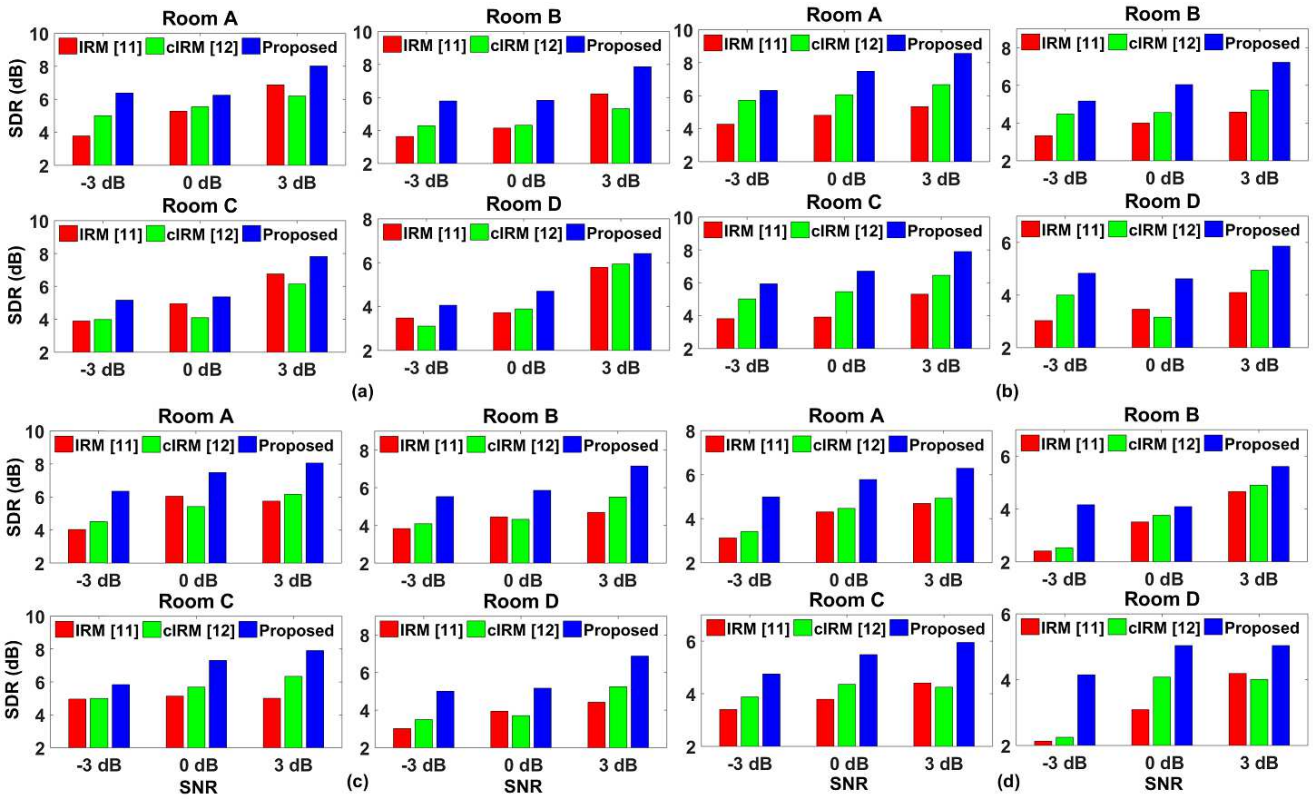

Fig. 2: The SDR (dB) in terms of different masks with various rooms. The X-axis is the SNR level, the Y-axis is the SDR (dB), each result is the average value of 120 experiments. The noise types in the (a), (b), (c) and (d) are factory, babble, cafe and SSN, respectively.

Tables 1 - 4 show the separation performance of the proposed method and the comparison groups with different noise in terms of the SNR_{fw} [18]. It is clear that the proposed method achieves the best performance in all scenarios and SNR levels. Although in some cases, such as the Room B with factory noise at -3 dB SNR level, the proposed method only improves the performance slightly when comparing with the cIRM, in other scenarios, the gains from the proposed method in terms of the SNR_{fw} are remarkable.

Moreover, when RT60 becomes higher, our proposed approach provides more significant improvements than the IRM and the cIRM. For example, in the Room A with babble noise, where the RT60 is the lowest, comparing with the cIRM, the further average SNR_{fw} improvements of the proposed method is 0.54 dB. While in the Room D with babble noise, where the RT60 is the highest, comparing with the cIRM, the corresponding average SNR_{fw} improvement is 1.16 dB. It can be observed that the proposed method is more efficient with high RT60s.

Figure 2 gives the separation performance in terms of SDR. It can be observed from the figure that the proposed method provides the largest SDR value consistently in all the scenarios. For example, in Figure 2(a), Room A, where the noise type is factory, the proposed method achieves 33.8 %, 37.4 % and 25.6 % improvements over the cIRM at -3, 0 and 3 dB SNR levels, respectively. In highly reverberant room environment, such as in Figure 2(d), Room D, where the RT60 is 0.89 s, the proposed method achieves 84.4 %, 57.8 % and 26.6 % improvements over the cIRM at -3, 0 and 3 dB SNR levels, respectively.

Comparing the performance with different RT60s, except for Room C, which has the largest direct-to-reverberant ratio (DRR) in these rooms, when the RT60 increases, the value of the SDR decreases. The separation performance improves with the increase in the SNR levels (from -3 dB to 3 dB).

In summary, by using the proposed IEM as the training target, the trained DNN model can generate a more effective T-F mask for source separation from the convolutive mixture. Compared with the IRM- and the cIRM-based DNN approaches, our proposed method provides the best performance in terms of SNR_{fw} and SDR consistently.

4. CONCLUSIONS AND FUTURE WORK

In this work, a dereverberation mask was firstly proposed and integrated with the IRM to generate the IEM. We demonstrated that the proposed method with the IEM as the training target gave better separation performance as compared with other masking-based methods, in terms of SNR_{fw} and SDR evaluations using different types of real RIRs and noise.

ISBN 978-90-827970-1-5 © EURASIP 2018

In the future work, we will try to solve the unseen RIRs problem and improve the generalization ability of the proposed method by using the advanced neural network architectures such as deep recurrent neural network (DRNN) or long short-term memory (LSTM) RNN.

5. REFERENCES

[1] Y. Sun, L. Zhu, J. A. Chambers, and S. M. Naqvi, "Monaural source separation based on adaptive discriminative criterion in neural networks," in *Proc. of IEEE International Conference on Digital Signal Processing (DSP)*, 2017.

[2] Y. Wang, K. Han, and D. L. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 270–279, 2013.

[3] M. S. Salman, S. M. Naqvi, A. Rehman, W. Wang, and J. A. Chambers, "Video-aided model-based source separation in real reverberant rooms," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1900–1912, 2013.

[4] B. Wu, K. Li, M. Yang, and C.-H. Lee, "A reverberation time aware approach to speech dereverberation based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 102–111, 2017.

[5] M. Yu, A. Rhuma, S. M. Naqvi, L. Wang, and J. A. Chambers, "A posture recognition-based fall detection system for monitoring an elderly person in a smart home environment," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 6, pp. 1274–1286, 2012.

[6] B. Rivet, W. Wang, S. M. Naqvi, and J. A. Chambers, "Audiovisual speech source separation: An overview of key methodologies," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 125–134, 2014.

[7] Y. Sun, W. Rafique, J. A. Chambers, and S. M. Naqvi, "Underdetermined source separation using time-frequency masks and an adaptive combined gaussian-students t probabilistic model," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[8] D. S. Williamson and D. L. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1492–1501, 2017.

[9] Z. Jin and D. L. Wang, "A supervised learning approach to monaural segregation of reverberation speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 625–638, 2009.

[10] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," *Speech Sep. Humans Mach*, vol. 60, pp. 63–64, 2005.

[11] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[12] D. S. Williamson, Y. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.

[13] IEEE Audio and Electroacoustics Group, "IEEE recommended practice for speech quality measurements," *IEEE Transactions on Audio Electroacoust*, vol. 17, no. 3, pp. 225–246, 1969.

[14] 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*, 1993.

[15] A. Varga and H. Steeneken, "Assessment for automatic speech recognition NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, pp. 247–251, 1993.

[16] C. Hummersone, R. Mason, and T. Brookes, "Dynamic precedence effect modeling for source separation in reverberant environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1867–1871, 2010.

[17] S.-H. Jin and C. Liu, "English sentence recognition in speech-shaped noise and multi-talker babble for English-, Chinese-, and Korean-native listeners," *Journal of the Acoustical Society of America*, vol. 132, no. 5, pp. 391–397, 2012.

[18] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[19] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transanctions on Audio Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

ISBN 978-90-827970-1-5