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

A feature study for masking-based reverberant speech separation

Masood Delfarah[1], *DeLiang Wang*[1,2]

[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Center for Cognitive and Brain Sciences, The Ohio State University, USA
delfarah.1@osu.edu, dwang@cse.ohio-state.edu

Abstract

Monaural speech separation in reverberant conditions is very challenging. In masking-based separation, features extracted from speech mixtures are employed to predict a time-frequency mask. Robust feature extraction is crucial for the performance of supervised speech separation in adverse acoustic environments. Using objective speech intelligibility as the metric, we investigate a wide variety of monaural features in low signal-to-noise ratios and moderate to high reverberation. Deep neural networks are employed as the learning machine in our feature investigation. We find considerable performance gain using a contextual window in reverberant speech processing, likely due to temporal structure of reverberation. In addition, we systematically evaluate feature combinations. In unmatched noise and reverberation conditions, the resulting feature set from this study substantially outperforms previously employed sets for speech separation in anechoic conditions.

Index Terms: speech separation, deep learning, room reverberation, feature selection, speech intelligibility

1. Introduction

Monaural speech separation refers to separating a target speaker from background interference in monaural recordings. Speech separation facilitates automatic speech recognition (ASR), and speaker identification (SID), enhances communication systems, and benefits hearing aid design. In this work, we employ deep neural networks (DNNs) to explore features for supervised speech separation in reverberant and noisy conditions, with the goal of improving human speech intelligibility.

Background noise presents a challenge to human listeners, since it overlaps with the target speech in time and frequency so that the speech is either rendered inaudible (i. e., energetic masking), or distorted in a way that the listener cannot segregate it from the background (i. e., informational masking). Room reverberation escalates the separation problem to a more demanding level and can be quite debilitating to human listeners. Several studies indicate that intelligibility scores of normal-hearing and hearing-impaired listeners drop significantly in the presence of background noise and room reverberation [1, 2]. Although deep learning algorithms have led to significant improvements of speech intelligibility in anechoic conditions [3], no intelligibility improvement has been achieved by monaural speech separation in noisy-reverberant conditions.

Time-frequency (T-F) masking is an effective separation approach introduced in computational auditory scene analysis (CASA) [4]. A supervised learning machine can be utilized to learn a mapping function from extracted acoustic features to an ideal T-F mask. The choice of a function approximation model and acoustic features is crucial for high-quality mask estimation. Incorporating a set of features can boost separation performance, as each feature may leverage some characteristics of the speech signal.

Feature combination for speech separation in anechoic conditions was first studied by Wang *et al.* [5] and later systematically evaluated by Chen *et al.* [6]. These studies, however, do not consider room reverberation which is unavoidable in realistic conditions. Therefore, it is questionable whether or not these feature sets are optimal in reverberant conditions. In addition, they draw their feature sets from group Lasso [7]. In Section 4.5 we will show that this method fails to consider generalization to new noises and room impulse responses (RIRs). Finally, we use DNNs as the supervised learning machine, which is shown to be more powerful than multilayer perceptrons and support vector machines used in previous studies. The investigation in this study results in a feature set that significantly outperforms the feature sets developed previously in noisy-reverberant conditions.

This paper is organized as follows. The evaluation framework is described in Section 2. Section 3 describes features to be investigated. In Section 4, we evaluate the effects of contextual information on separation performance, present evaluation results for individual features, and discuss feature combination. We conclude in Section 5.

2. Evaluation framework

The computational objective of masking-based speech separation is to estimate an ideal T-F mask obtained from premixed target and noise signals. Speech signal $s(t)$ and noise signal $n(t)$ sampled at 16 kHz are divided into 20 ms frames with 10 ms frame shift. Applying short-time Fourier transform (STFT) to each time frame results in 161 frequency bins. In this study, speech separation is formulated as estimating the ideal ratio mask (IRM) [8], defined as follows:

$$IRM(m,c) = \sqrt{\frac{S^2(m,c)}{S^2(m,c)+N^2(m,c)}} \tag{1}$$

where $S^2(.)$ and $N^2(.)$ represent the energy of the reverberant speech signal and diffuse noise energy at time frame m and frequency bin c.

One may consider the late or even early reverberation of the target speech as interference. Here, we choose to predict a mask for the magnitude spectrum of the entire reverberant target utterance. Studies suggest that human speech intelligibility does not drop significantly in reverberant but noiseless conditions [2]. Consequently, the IRM of Eq. 1 is expected to produce highly intelligible speech.

In order to obtain a fair comparison, we use a fixed DNN for IRM estimation. In the experiments, the DNN has three hidden layers with 512 hidden units in each layer. Activation

http://dx.doi.org/10.21437/Interspeech.2016-382

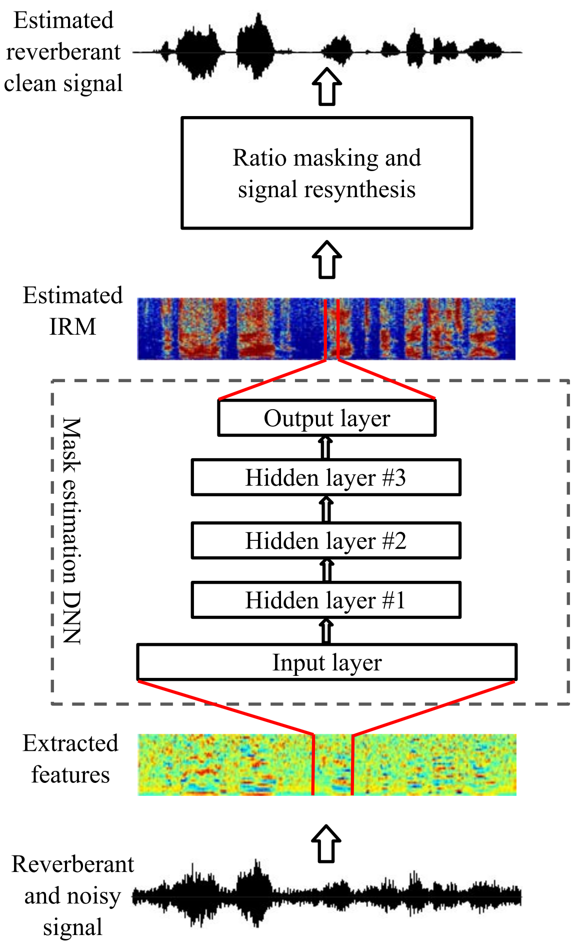

Figure 1: Diagram of the evaluation framework.

function for hidden units is the rectified linear function (ReLU) [9]. The sigmoid function is used as the output layer activation function. The backpropagation algorithm with gradient descent training is run for 25 epochs. The optimization objective is the following mean square error loss function:

$$\mathcal{L}(\boldsymbol{IRM}(m,:),\boldsymbol{F}(m);\Theta)=\sum_{c=1}^{C}(IRM(m,c)-g_c(\boldsymbol{F}(m)))^2 \qquad (2)$$

where $\boldsymbol{F}(.)$ denotes the feature vector, Θ corresponds to the model parameters, $C = 161$ is the number of frequency channels, and $g_c(.)$ is the output of the c^{th} output unit.

An overview of the evaluation framework is depicted in Fig. 1. Normalized features of a noisy and reverberant signal are passed through the trained DNN to obtain an estimated IRM. The obtained mask is applied to the magnitude spectrum of the noisy-reverberant speech to get the estimated clean magnitude spectrum. Finally, estimated clean signal is resynthesized from the noisy-reverberant signal phase and estimated clean magnitude spectrum.

3. Features

In this study, we evaluate a wide range of promising features utilized in anechoic and reverberant speech processing tasks. These features can be roughly categorized as gammatone-domain, spectral, or pitch-based. In the following, we briefly describe each of the features to be examined.

3.1. Gammatone-domain features

The noisy and reverberant signal is passed through a 64-channel gammatone filterbank [4], and the filter responses are decimated to the effective sampling rate of 100 Hz (equal to 10 ms frame shift). A cubic root operation is applied to the magnitude of decimated subband signals, resulting in 64-D gammatone feature (GF), per time frame.

A DCT operation on GF produces 31-D gammatone frequency cepstral coefficients (GFCC) [10]. With promising performance in speech separation in the anechoic environment, multi-resolution cochleagram feature (MRCG) introduced in [6] is designed to utilize local as well as contextual speech information. We follow the standard procedure to extract cochleagrams at four different resolutions to form a 256-D feature vector. Gammatone frequency modulation coefficient (GFMC) [11] is a recently developed reverberation-robust ASR feature. To extract GFMC, for each coefficient in GFCC, we calculate Fourier transform of the GFCC response, using 160 ms frame length and 10 ms frame shift. Energies for the 2-16 Hz modulation frequencies are calculated to generate a 31-D GFMC feature.

3.2. Pitch-based features

Pitch is an important cue in human speech perception, and has been employed in speech separation [12]. We use PEFAC [13] for pitch tracking, and then we calculate a 6-D feature described in [5] for each of the 64 gammatone filter outputs, to get a 384-D pitch-based feature (PITCH).

3.3. Spectral features

A 320-point STFT is applied to the input signal divided into 20 ms time frames with 10 ms frame shift, resulting in 161 Frequency bins of magnitude and phase responses. The magnitude responses are log-compressed to form 161-D log-spectral magnitude feature (LOG-MAG).

Various features have been proposed based on the Fourier transformation. We use the RASTAMAT toolbox [14] to extract three popular features including 31-D mel-frequency cepstral coefficients (MFCC), 13-D perceptual linear prediction (PLP), and 13-D relative spectral transformation PLP (RASTA-PLP) features.

Several studies have attempted to modify MFCC to obtain more noise- and distortion-robust features. Relative autocorrelation MFCC (RAS-MFCC) [15] applies the MFCC procedure to the high-pass filtered autocorrelation sequences, while phase autocorrelation MFCC (PAC-MFCC) [16] applies it to the phase angle between the signal and its shifted version. On the other hand, autocorrelation MFCC (AC-MFCC) [17] computes the autocorrelation of the signal in each frame, discards the low-lag coefficients, applies a window function, and finally, follows the MFCC procedure. In addition, we utilize suppression of slowly-varying components and the falling edge of the power envelope (SSF) [18] which is designed to enhance MFCC for robust ASR in reverberant conditions. We also use power normalized cepstral coefficients (PNCC) [19]. All these MFCC variants have the same dimensionality as the traditional MFCC.

We also evaluate 311-D Gabor filterbank (GFB) [20], and 15-D amplitude modulation spectrogram (AMS) [21] features.

4. Experimental results

4.1. Experiment setting

From the 720 male utterances in the IEEE corpus [22], 300 utterances are randomly chosen for training and validation, and the rest for testing. For background noise, cockpit, tank, factory, engine, vehicle, and speech-shaped noises from the NOISEX corpus [23] are used to corrupt the target utterances. We also generate an 8-talker babble noise, by mixing four male and four female randomly chosen speakers from the TIMIT corpus [24]. Factory, engine, and vehicle noises are used in the testing data in the unmatched noise evaluation. For the matched noise condition, the first half of each noise signal is used for training and the second half for testing.

The reverberant utterances at three different reverberation times (T_{60}) of 0.3, 0.6, and 0.9 s are created using the image method [25], introduced in [26]. We fix the simulation room size to (7,8,10) meters and place the microphone at the position of (3,4,1.5) meters and the speaker at a random place with 1 m distance from the microphone.

Diffuse babble noise is created by a symmetric placement of eight speakers at 2 m distance from the microphone. For all other six noise types, four random segments of each noise are placed at four different locations at 2 m distance from the microphone.

Each training mixture is created by mixing one IEEE utterance and one noise at one of the four signal-to-noise ratios (SNRs) of -9, -6, -3, and 0 dB. We also create anechoic training data (i. e., $T_{60} = 0.0$ s) at the aforementioned SNR levels. For each condition, we create 1000 training mixtures. Consequently, there are 4 (T_{60}s) $\times 4$ (SNRs) $\times 4$ (noise types) $\times 1000 = 64000$ training mixtures.

All of the test mixtures are generated at -6 dB SNR. Simulated room test data are generated following the same procedure as in training. For real room conditions, we use the recorded RIRs obtained from the 4 rooms in [27]. Note that no RIR, noise segment, or speech utterance are shared between the training and test mixtures.

4.2. Evaluation criterion

In this study, we employ short-time objective intelligibility (STOI) [28] which is a standard objective metric for intelligibility assessment. STOI score is a number in the range of 0 and 1, and higher values indicate higher human speech intelligibility scores.

We set the reference signal to the clean (noiseless) reverberant speech, and present the results in the form of percent STOI improvement as follows:

$$\Delta\mathrm{STOI}(\%) = 100 \times (\mathrm{STOI}_{\mathrm{processed}} - \mathrm{STOI}_{\mathrm{mixture}}) \quad (3)$$

4.3. Contextual effects

In order to capture temporal dynamics of reverberation we propose to incorporate features from the adjacent frames as:

$$\boldsymbol{F}_{a,b}(m) = [\boldsymbol{F}(m-a), \cdots, \boldsymbol{F}(m), \cdots, \boldsymbol{F}(m+b)] \quad (4)$$

where $\boldsymbol{F}(m)$ is the feature vector at time frame m.

Our experiments show that using contextual information leads to better separation results. Fig. 2 shows the effects of the contextual window size for GF on matched and unmatched test data in simulated room conditions. As seen in the figure, performance gain from using proceeding and succeeding frames is significant. In all of the following experiments we set $a = 3$, and $b = 3$, as a trade-off between computation cost and separation performance.

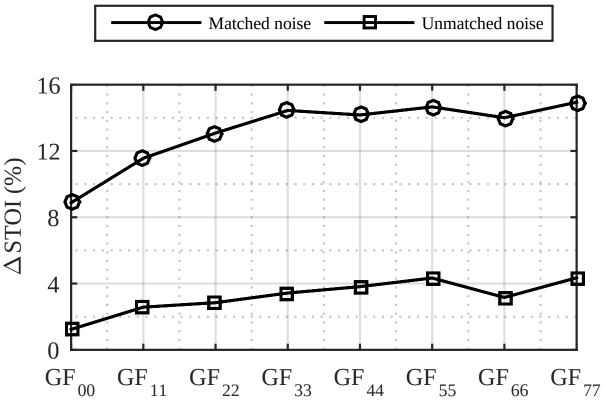

Figure 2: Average STOI improvement using GF with different contextual window sizes.

4.4. Single features

Table 1 gives the average ΔSTOI scores for each individual feature in matched noise conditions. The results are presented for anechoic, simulated, and real room conditions, and features are sorted in decreasing order of average performance. In general, gammatone-domain features have better scores in comparison to the others. MRCG is the best feature for anechoic and simulated room conditions; interestingly, PNCC outperforms MRCG in real room conditions.

Table 1: Feature performances in matched noise conditions. Average Δ STOI scores are presented. ”Ane”, ”Sim”, and ”Rec” indicate anechoic, simulated, recorded room condition, respectively. ”Avg” indicates average performance. The numbers in bold show the best results.

Feature	Ane	Sim	Rec	Avg
MRCG	**10.9**	**17.6**	11.8	**13.8**
PNCC	7.8	14.2	**13.4**	13.0
GF	9.9	16.0	11.4	12.9
GFCC	9.1	15.5	11.8	12.8
MFCC	9.3	15.0	10.7	12.1
LOG-MAG	7.6	14.5	11.3	12.0
PLP	8.4	13.9	9.7	11.1
RAS-MFCC	5.4	11.8	9.8	10.0
AC-MFCC	6.9	12.3	8.6	9.8
GFB	6.3	12.8	7.4	9.3
RASTA-PLP	4.5	10.0	9.5	9.1
SSF	4.1	9.4	7.8	8.0
PITCH	4.4	9.2	6.1	7.0
GFMC	4.0	9.1	4.8	6.3
PAC-MFCC	-0.5	-0.7	-0.8	-0.7
AMS	-4.9	-1.3	-7.2	-4.7

We present results for unseen noise conditions in Table 2 to investigate generalization capability. The results indicate that PNCC has significantly higher scores than other features. It is also interesting to note that in unseen noise conditions, except for MRCG, other gammatone-domain features do not clearly outperform other features like LOG-MAG and RAS-MFCC as in anechoic conditions [6].

Table 2: Feature performances in unmatched noise conditions. Average ΔSTOI scores are presented.

Feature	Ane	Sim	Rec	Avg
PNCC	**5.8**	**10.1**	**9.0**	**9.0**
MRCG	3.2	6.2	6.0	5.7
GFB	1.9	6.8	2.6	4.1
GFCC	1.8	4.1	3.3	3.4
GF	1.7	4.0	3.1	3.3
LOG-MAG	-0.2	4.5	3.3	3.3
RAS-MFCC	1.9	3.8	3.2	3.3
RASTA-PLP	1.3	3.6	3.2	3.1
SSF	1.7	3.3	2.8	2.9
PITCH	-1.3	3.6	3.0	2.7
MFCC	1.5	2.6	2.0	2.2
PLP	0.9	2.3	1.6	1.8
GFMC	0.0	1.5	0.6	0.8
AC-MFCC	0.4	1.6	0.2	0.7
PAC-MFCC	-2.4	-6.8	-3.2	-4.4
AMS	-6.5	-4.6	-7.7	-6.4

4.5. Feature combination

As seen in Section 4.4, features seem to exhibit characteristics in matched and unmatched conditions. In order to achieve further noise and reverberation robustness, in this subsection, we study feature combination.

The number of all possible feature combinations grows exponentially with respect to the number of individual features, and trying all such combinations is not feasible. The studies in [5] and [6] apply group Lasso [7] to identify complimentary feature sets in anechoic conditions. Group Lasso solves a group variable selection problem by introducing a mixed-norm regularization in linear regression. In [5], AMS+RASTA-PLP+MFCC ends up as a complementary feature set, while Chen *et al.* [6] suggest MRCG+PITCH as the combined features. We perform group Lasso on a development set from our train data, following the same procedure in [6]. Fig. 3 shows the average magnitude of regression coefficients across frequency channels. Significantly large regression coefficients in GFB, MRCG, LOG-MAG, PITCH features suggest complementarity in this group of features.

Group Lasso selects features based on a linear regression model. Conclusions on the basis of observations from a linear regression method may not apply to a highly nonlinear problem of speech separation. Furthermore, group Lasso does not take generalization into account, since it is agnostic to matched and unmatched conditions.

We employ a sequential floating forward selection (SFFS) algorithm [29] to determine the best feature set among the existing features. The SFFS algorithm starts with an empty set and repeatedly selects/drops features based on their relative performance, until no further gain is achieved. We apply SFFS to our development set, and based on the results, we propose the best *feature combination set* as GF+PNCC. Unlike the feature set resulted from the group Lasso approach, our proposed combination consists of GF, which performs well in matched conditions, and PNCC, a high-quality feature in unseen noises and room conditions.

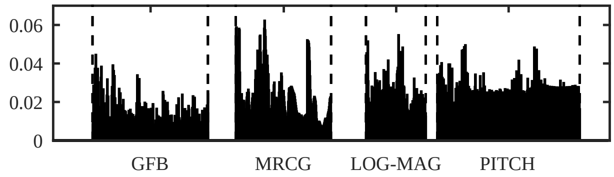

Figure 3: Normalized magnitude of coefficients from group Lasso.

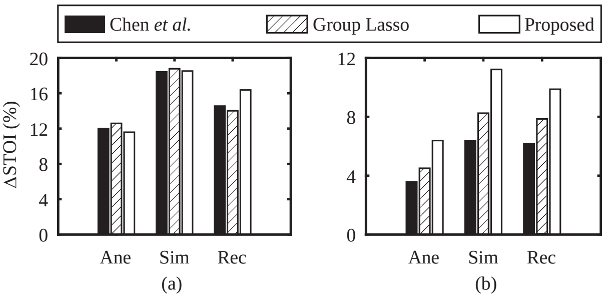

Figure 4: Average ΔSTOI for feature combination in matched noise conditions (a), and unmatched noise conditions (b).

We present our feature combination performance, and compare it with the feature set suggested by group Lasso, and the feature set introduced in Chen *et al.* [6]. Fig. 4 shows average ΔSTOI scores in matched and unmatched noise conditions. As shown in the figure, in the anechoic and simulated room conditions with matched noises, the three feature sets have similar performances. It is important to note that the dimensionality of the proposed feature set is 665 $(7 \text{ frames } \times(64 + 31))$, while in the feature sets from group Lasso and Chen *et al.* [6] this number is 7784 and 4480, respectively. As a result, our feature set has a significantly lower computational cost.

In unmatched noise conditions, the proposed feature set substantially outperforms the other two, indicating its generalization power to new acoustic conditions. Comparing the results in Fig. 4, and the individual feature scores in Tables 1 and 2, the proposed feature set outperforms the features in all of the conditions. Unlike the individual features, the proposed feature set is consistently effective in matched and unmatched noises and different room conditions.

5. Conclusion

In this paper, we have conducted a feature study for masking-based speech separation in reverberant conditions at a very low SNR level, and different room conditions and noise types. We have also proposed a feature combination that, based on objective speech intelligibility, performs significantly better than previous feature sets developed for anechoic conditions. In future research, we will extend the present work to cochannel speech (i. e., two-talker) separation.

6. Acknowledgements

We thank Jitong Chen for providing help with feature extraction. This research was supported in part by an AFOSR grant (FA9550-12-1-0130) and the Ohio Supercomputer Center.

7. References

[1] K. S. Helfer and L. A. Wilber, "Hearing loss, aging, and speech perception in reverberation and noise," *Journal of Speech, Language, and Hearing Research*, vol. 33, pp. 149–155, 1990.

[2] E. L. George, S. T. Goverts, J. M. Festen, and T. Houtgast, "Measuring the effects of reverberation and noise on sentence intelligibility for hearing-impaired listeners," *Journal of Speech, Language, and Hearing Research*, vol. 53, pp. 1429–1439, 2010.

[3] E. W. Healy, S. E. Yoho, Y. Wang, and D. L. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 134, pp. 3029–3038, 2013.

[4] D. L. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE Press, 2006.

[5] Y. Wang, K. Han, and D. L. Wang, "Exploring monaural features for classification-based speech segregation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, pp. 270–279, 2013.

[6] J. Chen, Y. Wang, and D. L. Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, pp. 1993–2002, 2014.

[7] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, pp. 49–67, 2006.

[8] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, pp. 1849–1858, 2014.

[9] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *International Conference on Artificial Intelligence and Statistics*, 2011, pp. 315–323.

[10] Y. Shao and D. L. Wang, "Robust speaker identification using auditory features and computational auditory scene analysis," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 1589–1592.

[11] H. K. Maganti and M. Matassoni, "An auditory based modulation spectral feature for reverberant speech recognition," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[12] H. Zhang, X. Zhang, S. Nie, G. Gao, and W. Liu, "A pairwise algorithm for pitch estimation and speech separation using deep stacking network," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 246–250.

[13] S. Gonzalez and M. Brookes, "A pitch estimation filter robust to high levels of noise (PEFAC)," in *Signal Processing Conference, 2011 19th European*. IEEE, 2011, pp. 451–455.

[14] D. Ellis, "PLP and RASTA (and MFCC, and inversion) in matlab using melfcc. m and invmelfcc. m," 2006. [Online]. Available: http://labrosa.ee.columbia.edu/matlab/rastamat/

[15] K.-H. Yuo and H.-C. Wang, "Robust features for noisy speech recognition based on temporal trajectory filtering of short-time autocorrelation sequences," *Speech Communication*, vol. 28, pp. 13–24, 1999.

[16] S. Ikbal, H. Misra, and H. Bourlard, "Phase autocorrelation (PAC) derived robust speech features," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 2. IEEE, 2003, pp. II–133.

[17] B. J. Shannon and K. K. Paliwal, "Feature extraction from higher-lag autocorrelation coefficients for robust speech recognition," *Speech Communication*, vol. 48, pp. 1458–1485, 2006.

[18] C. Kim and R. M. Stern, "Nonlinear enhancement of onset for robust speech recognition." in *INTERSPEECH*, 2010, pp. 2058–2061.

[19] C. Kim and R. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4101–4104.

[20] M. R. Schädler, B. T. Meyer, and B. Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 131, pp. 4134–4151, 2012.

[21] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *The Journal of the Acoustical Society of America*, vol. 126, pp. 1486–1494, 2009.

[22] E. Rothauser, W. Chapman, N. Guttman, K. Nordby, H. Silbiger, G. Urbanek, and M. Weinstock, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust*, vol. 17, pp. 225–246, 1969.

[23] 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, pp. 247–251, 1993.

[24] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, 1993.

[25] E. A. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, p. 1, 2006.

[26] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, pp. 943–950, 1979.

[27] C. Hummersone, R. Mason, and T. Brookes, "Dynamic precedence effect modeling for source separation in reverberant environments," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, pp. 1867–1871, 2010.

[28] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, pp. 2125–2136, 2011.

[29] P. Pudil, F. Ferri, J. Novovicova, and J. Kittler, "Floating search methods for feature selection with nonmonotonic criterion functions," in *Proceedings of the Twelveth International Conference on Pattern Recognition, IAPR*. Citeseer, 1994.