Interspeech 2018
2-6 September 2018, Hyderabad

A Deep Neural Network Based Harmonic Noise Model for Speech Enhancement

Zhiheng Ouyang[1], *Hongjiang Yu*[1], *Wei-Ping Zhu*[1], *Benoit Champagne*[2]

[1]Dept. of Electrical and Computer Engineering, Concordia University, Montreal, Canada
[2]Dept. of Electrical and Computer Engineering, McGill University, Montreal, Canada

z_ouyan@encs.concordia.ca, ho_yu@encs.concordia.ca, weiping@ece.concordia.ca, benoit.champagne@mcgill.ca

Abstract

In this paper, we present a novel deep neural network (DNN) based speech enhancement method that uses a harmonic noise model (HNM) to estimate the clean speech. By utilizing HNM to model the clean speech in the short-time Fourier transform domain and extracting some time-frequency features of noisy speech for the DNN training, the new method predicts the harmonic and residual amplitudes of clean speech from a set of noisy speech features. In order to emphasize the importance of the harmonic component and reduce the effect caused by the residual, a scaling factor is also introduced and applied to the residual amplitude. The enhanced speech is reconstructed with the estimated clean speech amplitude and the noisy phase of HNM. Experimental results demonstrate that our proposed HNM-DNN method outperforms two existing DNN based speech enhancement methods in terms of both speech quality and intelligibility.

Index Terms: speech enhancement, deep neural network, harmonic noise model

1. Introduction

In real world environments, clean speech is often corrupted by a wide range of background noises, which causes problems in applications including voice communication, automatic speech recognition and speaker identification. Speech enhancement, which aims to improve speech quality and intelligibility, has been intensively studied over the past decades to obtain better user experience in speech processing, recognition and communication.

Many traditional speech enhancement methods have been foused on estimating the short time spectral amplitude (STSA) of clean speech. Among them, Wiener filtering [1] and minimum mean square error (MMSE) amplitude estimators [2] are two most well-known techniques. In Wiener filtering approach, the estimated speech spectrum was obtained by multiplying a Wiener gain function to the noisy spectrum, where the Wiener gain was derived by minimizing the mean square error between the clean and estimated speech spectrums. The MMSE estimator of clean speech was derived by minimizing the statistical expectation of a cost function that penalizes the error in the clean speech estimation. These estimators achieve good results to some extent, but they also cause distortion of speech signals. In addition, due to inaccuracies in the estimation of speech and noise statistics, both Wiener filter and MMSE estimator suffer from residual noise which has an annoying noticeable effect on the enhanced speech.

Different from the above STSA estimator-based methods, which mainly focus on the enhancement of speech quality, time-frequency (TF) masking is a kind of approach that attempts to improve speech intelligibility. This masking technique amounts to selecting a subset of frequency bins from the corrupted speech spectra while discarding the rest. Ideal binary mask (IBM) [3] and ideal ratio mask (IRM) [4] are two well-known masking techniques in this area, and their performance depends largely on the quality of TF mask estimation.

Nowadays the deep neural network (DNN) based speech enhancement methods are getting more and more popular, as this kind of supervised methods have the potential to deal with more complex acoustic environment. In [5], a DNN framework was proposed to directly restore the clean speech amplitude by building a mapping function between the log spectral power (LPS) feature of the noisy speech and that of the clean speech. By training the DNN with a large set that encompasses many possible combinations of speech and noise types, the estimated speech has significantly better objective and subjective measures compared to that achieved by conventional MMSE-based techniques. While using DNN to directly estimate clean speech amplitude is intuitive, it requires a large training set in order to form a mapping as accurately as possible. In [6], the learning target of DNN was turned to estimating IRM, and then the enhanced speech was obtained by applying the estimated mask to noisy speech, resulting in a better denoising performance. This method has benefited from the masking effect, namely, it discarded the noise-dominant part and maintained the speech-dominant part in the mixed signals. On the other hand, although this method can improve speech intelligibility, it harms the underlying clean speech at the same time.

Many speech enhancement algorithms ignored the harmonic structure of speech spectrum, and thus suffered from a poor enhancement performance, especially in low signal-to-noise ratio (SNR) situations [7]. It is well-known that using HNM for modeling the harmonic structure yeilds better intelligibility and quality of synthesized speech. In [8], the authors proposed a restoration method to retrieve the HNM of original speech from the damaged version, in which the estimated parameters of HNM are obtained by utilizing a pre-trained codebook. In [9], an estimator based on HNM was derived for speech separation. The proposed estimator aimed to find the HNM parameters from the pre-trained speakers' codebooks. Then, the separated speech is obtained by applying the estimator to mixed signals. Compared to binary mask based separation algorithms [3], this method achieved a better perceptual speech quality.

In this paper, we propose for the first time a DNN based HNM for noise reduction of speech signals in the frequency domain by exploring the relationship between the HNM parameters of noisy speech and that of clean one. In contrast to the codebook techniques for estimating HNM parameters described

10.21437/Interspeech.2018-1114

above, our approach makes use of DNN's ability of learning complex non-linear mapping function, that is, using DNN to learn the HNM parameters of clean speech from the spectrum features of noisy speech. In addition, the target of our algorithm is different from the above mentioned DNN based denoising methods. Previous methods either attempted to predict the clean speech amplitude or obtained an estimated mask, while our approach emphasizes on the harmonic structure of speech. As a consequence of better restoring the harmonic structure of speech signals, our algorithm gives better speech quality and intelligibility, especially in low SNR acoustic environments.

2. Harmonic Noise Model of Speech

HNM of speech was first introduced by Stylianous [10], which divided the speech into harmonic part and noise or residual part. In order to distinguish the noise part in HNM from the background noise, we will call it residual in the rest of the paper. Thus, the speech signal $s\left(n\right)$ can be written as

$$s\left(n\right)=h\left(n\right)+r\left(n\right) \tag{1}$$

where $h\left(n\right)$ and $r\left(n\right)$ represent the harmonic component and the residual of the speech, respectively.

The harmonic part of speech conveys most of the voiced speech information and can be modeled as a weighted superposition of a series of sinusoids at the fundamental frequency f_0 and its harmonic frequencies, i.e.,

$$h\left(n\right)=\sum_{i=1}^{I}a_{h,i}\left(n\right)\cos\left(\varphi_{h,i}\left(n\right)\right) \tag{2}$$

with

$$\varphi_{h,i}\left(n\right)=\Omega_i\left(n\right)n+\phi_i=2\pi n\frac{f_i}{f_0}+\phi_i \tag{3}$$

where I denotes the number of harmonics, which can be computed as $\lfloor f_s/2f_0 \rfloor$, with f_s being the sampling frequency and $\lfloor\ \rfloor$ the floor operator, $a_{h,i}$ and $\varphi_{h,i}$ are the time-domain amplitude and phase of the i-th harmonic component, Ω_i the normalized angular frequency and $f_i=\left(i+1\right)f_0$ the harmonic frequencies. Finally, ϕ_i is the initial time domain phase of the i-th harmonic component.

The residual of speech is obtained by subtracting the harmonic part from the original speech signal in the time-domain. For clean speech $s_x(n)$, the residual usually accounts for the non-periodic components of speech signal, such as fricative or aspiration noise, periodic variations of the glottal excitation and so on; while for noisy speech $s_y(n)$ the residual contains both non-periodic speech and additive background noise.

The time domain harmonic noise model of speech can be transformed into short-time discrete Fourier transform (STFT) domain. The corresponding STFT representation of (1) is denoted as

$$S(l,k)=H(l,k)+R(l,k) \tag{4}$$

with frame index l and frequency index k. In order to simplify the equations, we will omit l in the remaining discussion. We denote the complex spectral coefficients of $s(n)$, $h(n)$, and $r(n)$ by the corresponding capital letters which can be described in terms of their amplitudes and phases, namely,

$$\begin{aligned}S\left(k\right)&=A_s\left(k\right)e^{j\Phi_s\left(k\right)}\\H\left(k\right)&=A_h\left(k\right)e^{j\Phi_h\left(k\right)}\\R\left(k\right)&=A_r\left(k\right)e^{j\Phi_r\left(k\right)}\end{aligned} \tag{5}$$

Our goal in this study is to estimate the amplitude of HNM of clean speech in the frequency domain, including both the harmonic amplitude $A_h(k)$ and residual amplitude $A_r(k)$. The two estimated amplitudes will be combined together with the phase of the noisy speech to reconstruct the enhanced speech. The reason to use the noisy phase is because human ear is less sensitive to the changes of the phase [11]. Although the authors of [12] showed that using the clean phase for the reconstruction can improve the quality of the enhanced speech, yet it is difficult to predict the clean phase in practical applications due to the randomness property of the phase. Hence in this paper we only take the amplitude estimation of HNM into consideration.

3. Proposed Speech Enhancement System

The overall block diagram of our speech enhancement system is depicted in Fig.1. A DNN is employed to find a mapping between the time-frequency features of noisy speech $s_y(n)$ and the amplitude of HNM of clean speech $s_x(n)$. Our system consists of two stages: training stage and enhancement stage. In the training stage, we extract noisy speech features as the input of DNN, and use the clean amplitude of HNM as the target of DNN. Then DNN is trained to minimize the difference between the estimated amplitude and the clean amplitude of HNM. In the enhancement stage, the noisy speech features are extracted and processed by the well trained DNN to predict the clean amplitude of HNM. The enhanced speech is then synthesised using the estimated clean amplitude of HNM and the noisy phase.

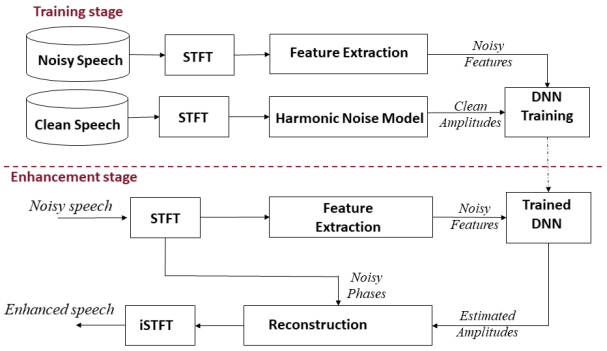

Figure 1: *A bolck diagram of proposed speech enhancement system.*

3.1. Speech Features

An ideal feature set is capable of predicting the target of DNN accurately. A complementary set of features was analysed in [13]. They are amplitude modulation spectrum (AMS), relative spectral transform and perceptual linear prediction (RASTA-PLP), Mel-frequency cepstral coefficients (MFCC) as well as their deltas, and Gammatone filterbank (GF) energies as well as their deltas. These features are computed for each frame of the signal and expected to successfully perform speech separation tasks.

To make full use of the temporal information of speech, it is a common way to incorporate features of adjacent time frames into a single feature vector. Hence, the feature vector centered at the l-th frame is constructed as $\tilde{\mathbf{F}}\left(l\right)=\left[\mathbf{F}\left(l-p\right),\ldots,\mathbf{F}\left(l\right),\ldots,\mathbf{F}\left(l+p\right)\right]$, where p represents the number of adjacent frames to be appended on each side.

3.2. Training Target

Our training target includes harmonic amplitude A_{hx} and residual amplitude A_{rx} of clean speech in the frequency domain. The harmonic amplitude A_{hx} is computed as

$$A_{hx}(k) = \begin{cases} A_{sx}(k), & k = \lfloor if_0N/f_s \rfloor \\ 0, & otherwise \end{cases} \quad (6)$$

where A_{sx} is the amplitude of the STFT of clean speech , N is the length of discrete Fourier transform (DFT), $i \in \{1, 2, ..., I\}$ the index of harmonics.

We use pitch estimator with amplitude compression (PEFAC) [14] to obtain a good estimate of fundamental frequency f_0, as this method can estimate pitch reliably even at low SNRs. Then, residual amplitude A_{rx} can be obtained by subtracting A_{hx} from A_{sx} .

As we stated before, the harmonic part mainly accounts for the voiced component of the speech signal. So we attempt to concentrate more on enhancing the harmonic part. To this end, a scaling factor $\alpha \in [0, 1]$ will be applied to the residual amplitude. The residual amplitude remains unsuppressed if we set the scaling factor to 1, while it is totally discarded when the scaling factor is set to 0. From our experiments, the favourable scaling factor is given by

$$\alpha = \begin{cases} 1, & q > 0.85 \\ \sqrt{\frac{A_{rx} - A_{rx,\min}}{A_{rx,\max} - A_{rx,\min}}}, & otherwise \end{cases} \quad (7)$$

where $A_{rx,\min}$ is the minimal residual amplitude in each frame, $A_{rx,\max}$ is the maximum, and q is the ratio of the clean amplitude to the noisy amplitude. A situation where q is under a preset threshold implies that the speech is degraded by the background noise. On the contrary, if the ratio is above the threshold, we suppose the background noise has only a slight impact on the speech only, and the residual amplitude remains intact. In our paper, the threshold is set to 0.85 based on our large experimentation. Then, the scaled residual amplitude is given by $\tilde{A}_{rx} = \alpha A_{rx}$, which reduces the impact of the residual amplitude on the reconstruction of the enhanced speech, thus giving better enhancement result.

We use both unscaled target and scaled target in our experiments to demonstrate the effect of the residual. The unscaled target is $[A_{hx}, A_{rx}]$, i.e., the scaling factor is set to 1. We call this type as basic target. The scaled target is illustrated by $[A_{hx}, \tilde{A}_{rx}]$, which includes the full harmonic target as the special case when the scaling factor is set to 0.

3.3. DNN Based Estimation

As Fig.2 depicts, a DNN is trained to learn the mapping from noisy speech features to the amplitude of the clean speech HNM. Our DNN consists of one input layer, one output layer and three hidden layers with 512 units in each layer. The activation function used in the hidden layer is the rectified linear unit (ReLU), while a linear function is used in the output layer.

In this work, the aforementioned AMS, RASTA-PLP, MFCC and GF are used as a feature set. The input noisy speech feature vector has dimensionality of $R(2p + 1)$ with the combination of adjacent frames, where R is the dimension of the feature set of a single frame, which is set to 246, and p is set to 1 in our experiments.

Back propagation is used to train the DNN. The cost function is defined as the following total mean square error (MSE)

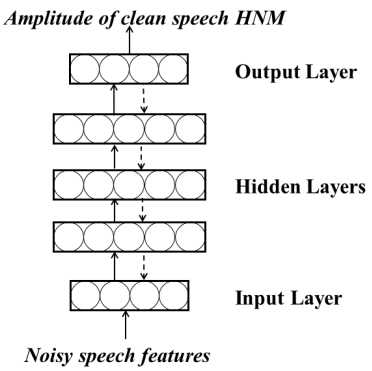

Figure 2: *Structure of the proposed DNN framework.*

over M speech frames,

$$\frac{1}{M}\sum_{m=1}^{M}\left[\sum_i \left(\hat{A}_{i,hx} - A_{i,hx}\right)^2 + \sum_j \left(\hat{A}_{j,rx} - A_{j,rx}\right)^2\right] \quad (8)$$

which measures the difference between the estimated harmonic and residual amplitude components and the corresponding clean speech components for the basic target. In (8), M denotes the number of frames, m the frame index, i the index of harmonics, j the index of residuals, $\hat{A}$ the estimated HNM amplitude, and A that of clean speech.

After obtaining the estimated amplitude from the DNN, the STFT of enhanced speech $\hat{S}_x(k)$ is reconstructed together with the noisy phases Φ_{sy}, i.e.,

$$\hat{S}_x(k) = \hat{A}_{sx}(k) e^{j\Phi_{sy}(k)} \quad (9)$$

with

$$\hat{A}_{sx}(k) = \hat{A}_{hx}(k) + \hat{A}_{rx}(k) \quad (10)$$

The enhanced speech $\hat{s}_x(n)$ can be obtained by performing the inverse STFT of $\hat{S}_x(k)$. It is important to note that although (8) and (9) only illustrate the situation where the DNN is trained with the basic target, the scaled target case can be processed.

4. Experimental Results

4.1. Experimental Setup

In this study, the clean speech is selected from the IEEE corpus [15]. We choose 300 utterances for training and 80 utterances for testing. Eight types of noises are picked form NOISEX-92 database [16]. Among them four types (babble, white, buccaneer, factory) are regarded as seen noises, and the others (pink, hfchannel, destroyerops, f16) as unseen noise. In the training stage, the noisy speech are obtained by mixing clean training speech with seen noises at four levels of SNRs (-10dB, -5dB, 0dB, 5dB), which results in 4800 utterances. In the testing stage, both seen noises and unseen noises are mixed with clean testing speech at the above four SNRs, so the number of noisy utterances used in enhancement stage is 1280 for seen noise and 1280 for unseen noise, respectively. A 16 kHz sampling rate is used for each signal. We use hamming window to divide each signal into 20 ms time frames with an 5 ms frame shift (i.e. 75% overlap). A 320-point DFT is then computed with each frame consisting of 161 samples.

We compare our speech enhancement algorithm with two existing DNN based methods: LPS-DNN [5] and IRM-DNN

[6]. For the proposed HNM-DNN method, three variations are considered, which correspond to different scaling factors used in the training stage: bHNM-DNN for the basic case with $\alpha = 1$, sHNM-DNN for the scaled case in (7) and HM-DNN for the full harmonic case with $\alpha = 0$.

Here, we adopt the perceptual evaluation of speech quality (PESQ) [17] and short-time objective intelligibility (STOI) [18] for objective assessments for the enhanced speech. PESQ focus on evlauating speech quality while STOI evaluating speech intelligibilty. In addition, the frequency-weighted segmental SNR (SSNR) [19] is also used.

4.2. Results and Discussions

Table 1 gives the average objective score of different DNN based speech enhancement algorithms on seen noises. Clearly, our proposed HNM-DNN method has better overall objective scores than the other two DNN based methods in most cases, except for the STOI score of the IRM-DNN in 5dB scenario. This is because the masking effect of IRM-DNN works well and enhances the speech intelligibility when the input SNR is high. However, our HNM-DNN shows a good performance on STOI scores, since the speech intelligibility is less likely to be degraded by the noise in high SNR environments. At low SNR levels, our HNM-DNN achieves better results in terms of all three metrics, because it emphasizes on restoring the harmonic structure, leading to a better enhancement for the voiced speech.

The result of HNM-DNN also differs when changing the value of the scaling factor. Firstly, HM-DNN has the best SSNR in low SNR cases, because it only aims to better restore voiced speech in the noisy speech. However, sHNM-DNN achieved the best SSNR at high SNRs, since the residual also contains unvoiced speech information in this case. Totally discard the residual will decrease the enhancement performance. Secondly, bHNM-DNN performs the best in terms of STOI metric, since its goal is to model the whole amplitude of the speech, which preserves the richest information in speech. On the other hand, sHNM-DNN also achieves similar STOI scores as bHNM-DNN, but the PESQ scores are much better. This indicates that the scaling factor computed by (7) can suppress the residual noise in the noisy speech. As a result, sHNM-DNN appears to have the best performance after considering all aspects of objective evaluation metrics.

Table 2 shows the average objective score of different DNN based speech enhancement algorithms on unseen noises. Our HNM-DNN method still obtains the best objective scores on the whole. It should be noticed that the LPS-DNN achieves high SSNR while the PESQ and STOI are not improved apparently. After subjective listening tests, we found that LPS-DNN removes speech and noise simultaneously in these cases. Although the SSNR is much higher, the PESQ and STOI scores are not satisfactory.

Table 1: *Speech enhancement results on seen noisy speeches*

		-10dB	-5dB	0dB	5dB
PESQ	Noisy	1.093	1.297	1.533	1.821
	LPS-DNN	0.871	1.356	1.905	2.369
	IRM-DNN	1.427	1.864	2.305	2.672
	HM-DNN	1.424	1.925	2.200	2.460
	bHNM-DNN	1.506	1.925	2.362	2.701
	sHNM-DNN	**1.578**	**2.012**	**2.445**	**2.765**
STOI	Noisy	0.498	0.586	0.684	0.785
	LPS-DNN	0.448	0.598	0.723	0.817
	IRM-DNN	0.580	0.715	0.827	**0.896**
	HM-DNN	0.585	0.707	0.791	0.837
	bHNM-DNN	**0.601**	**0.730**	**0.831**	0.890
	sHNM-DNN	0.591	0.724	0.825	0.884
SSNR	Noisy	-12.277	-10.046	-6.485	-2.092
	LPS-DNN	-2.067	-0.621	1.498	3.475
	IRM-DNN	-2.936	-0.009	2.778	5.435
	HM-DNN	**-1.660**	**0.643**	3.117	4.883
	bHNM-DNN	-2.462	-0.071	2.923	5.683
	sHNM-DNN	-2.186	0.270	**3.445**	**6.155**

Table 2: *Speech enhancement results on unseen noisy speeches*

		-10dB	-5dB	0dB	5dB
PESQ	Noisy	1.123	1.325	1.556	1.847
	LPS-DNN	1.021	1.522	2.019	2.445
	IRM-DNN	1.365	1.700	2.051	2.416
	HM-DNN	1.405	1.765	2.078	2.327
	bHNM-DNN	1.388	1.734	2.105	2.462
	sHNM-DNN	**1.442**	**1.801**	**2.167**	**2.512**
STOI	Noisy	0.516	0.601	0.699	0.798
	LPS-DNN	0.518	0.646	0.760	0.838
	IRM-DNN	0.553	0.672	0.783	0.868
	HM-DNN	**0.581**	0.697	0.781	0.834
	bHNM-DNN	0.580	**0.706**	**0.805**	**0.874**
	sHNM-DNN	0.569	0.696	0.798	0.868
SSNR	Noisy	-12.672	-10.237	-6.608	-2.129
	LPS-DNN	**-1.003**	**0.428**	**2.378**	3.981
	IRM-DNN	-8.374	-5.104	-1.198	2.824
	HM-DNN	-4.166	-0.967	2.057	**4.634**
	bHNM-DNN	-6.320	-3.066	0.381	3.894
	sHNM-DNN	-6.037	-2.673	1.067	4.587

5. Conclusions

In this paper, HNM-DNN based technique has been proposed for speech enhancement. Compared with the LPS-DNN and IRM-DNN techniques, our approach has aimed at recovering the harmonic structure of the speech, which results in superior speech quality as well as intelligibility. It is also found that by adding a scaling factor to the residual amplitude in the training stage, the speech quality can be further enhanced without degrading severely the speech intelligibility.

It should be noted that only the amplitude of clean speech HNM has been estimated in our method, while the phase remains the same as the noisy speech HNM. One could obtain a better performance by synthesizing the enhanced speech with the phase of the clean speech HNM. In that case, however, we need to make special effort to estimate the phase of the clean speech HNM, which obviously increases the computational complexity of the enhancement system.

6. Acknowledgements

The authors would like to acknowledge the support of China Scholarships Council (CSC No.201606270200) and the Natural Sciences and Engineering Research Council (NSERC) of Canada under a CRD project sponsored by Microsemi, Ottawa, Canada.

7. References

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

[2] 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 (TASLP)*, vol. 32, no. 6, pp. 1109–1121, 1984.

[3] D. Wang and G. J. Brown, "Computational auditory scene analysis: Principles, algorithms, and applications," 2006.

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

[5] 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 (TASLP)*, vol. 23, no. 1, pp. 7–19, 2015.

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

[7] C. Plapous, C. Marro, and P. Scalart, "Speech enhancement using harmonic regeneration," in *Acoustics, Speech and Signal Processing (ICASSP), 2005 IEEE International Conference on*. IEEE, 2005, pp. 157–160.

[8] E. Zavarehei, S. Vaseghi, and Q. Yan, "Noisy speech enhancement using harmonic-noise model and codebook-based post-processing," *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 15, no. 4, pp. 1194–1203, 2007.

[9] P. Mowlaee, M. G. Christensen, and S. H. Jensen, "New results on single-channel speech separation using sinusoidal modeling," *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 19, no. 5, pp. 1265–1277, 2011.

[10] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Transactions on speech and audio processing (TASLP)*, vol. 9, no. 1, pp. 21–29, 2001.

[11] E. A. Wan and A. T. Nelson, "Networks for speech enhancement," *Handbook of neural networks for speech processing. Artech House, Boston, USA*, vol. 139, p. 1, 1999.

[12] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 55–66, 2015.

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

[14] S. Gonzalez and M. Brookes, "Pefac-a pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 2, pp. 518–530, 2014.

[15] I. Subcommittee, "Ieee recommended practice for speech quality measurements," *IEEE Trans. on Audio and Electroacoustics*, vol. 17, pp. 225–246, 1969.

[16] 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.

[17] ITU-R, "Perceptual evaluation of speech quality (pesq) an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Recommendation P.862*, 2001.

[18] 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 (TASLP)*, vol. 19, no. 7, pp. 2125–2136, 2011.

[19] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *The Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3387–3405, 2009.