A Speech Reconstruction Algorithm via Iteratively Reweighted ℓ_2 Minimization for MFCC Codec

1st Gang Min
Institute of Information and Communication
National University of Defense Technology
Xi'an, China
mgxaty@gmail.com

2nd Xiongwei Zhang
Institute of Command and Control
Army Engineering University
Nanjing, China
xwzhang@gmail.com

3rd Xiangyang Liu
Institute of Information and Communication
National University of Defense Technology
Xi'an, China
liuxiangyang@foxmail.com

4th Changqing Zhang
Institute of Information and Communication
National University of Defense Technology
Xi'an, China
zhangcq1108@163.com

5th Yanpu Chen
Institute of Information Engineering
Xijing University
Xi'an, China
dzxxlab@163.com

***Abstract*—This paper presents an effective method to address the inverse problem of Mel-frequency cepstral analysis, and describes how to reconstruct the speech waveforms from Mel-frequency cepstral coefficients (MFCCs) directly. To exploit the sparse characteristics of speech in the frequency domain, an iteratively reweighted ℓ_2 minimization method is proposed to cope with the under-determined nature of the reconstruction problem. The lost phase information during Mel-frequency cepstral analysis procedure is recovered by the inverse short-time Fourier transform magnitude algorithm. Experiments are conducted over the TIMIT database and evaluated by several different kinds of measures. Experimental results demonstrate that the proposed method recovers speech with high articulation and intelligibility. Specifically, it sounds very close to the original speech when using the high-resolution MFCCs, the average STOI, PESQ score reaches 93% and 4.0, respectively. This method could be easily used for MFCC codec at low bit rate.**

***Index Terms*—Speech reconstruction, MFCCs, Iteratively reweighted ℓ_2 minimization**

I. Introduction

The Mel-frequency cepstral analysis of speech signals, which converts the speech waveforms to MFCCs, is an important homomorphic signal processing technique. MFCCs are widely used in the applications of speech formant analysis, automatic speech recognition and speaker recognition, etc [1]. In recent years, MFCC codec was proposed to encode the speech signal through quantization of MFCCs, which provides a promising new approach for speech coding throughout 600-4800 bps [2] [3]. In the MFCC codec, the challenging inverse problem of Mel-frequency cepstral analysis, i.e., reconstructing the speech waveforms from MFCCs, is a key step, which attracts growing attention of researchers [4].

Generally speaking, there are two difficult problems for reconstructing the speech waveforms from MFCCs. Firstly, the phase information is lost when computing the power spectrum of speech. In addition, the Mel-filter, i.e., the Mel-scale weighting matrix, is not invertible. Based on the sinusoidal model, techniques were proposed in [5] [6] to reconstruct the speech waveforms from MFCCs. However, additional pitch and voicing decision information are indispensable. These papers reported that "natural sounding, good quality intelligible speech" was obtained. Similarly, the ETSI standardized the extended DSR as ES 202 211 and ES 202 212 [7] [8]. Experimental results show that the reconstructed speech produced by these standards is highly intelligible under clean and noisy background conditions, the Diagnostic Rhyme Test (DRT) and Transcription Test (TT) scores meet or exceed the US Federal standard Mixed-Excitation Linear Predictive (MELP) codec operating at 2400 bps. In addition, a new algorithm for speech reconstruction solely from MFCC vectors was proposed in [9]. This algorithm predicts pitch frequency and voicing information by exploiting correlation between the fundamental frequency and the spectral envelope. The speech waveforms is then reconstructed with the method in DSR back-end, which is similar to [7] [8].

Different from the earlier investigations, the authors in [2], [4] proposed a simple and novel speech reconstruction method. In this novel method, the power spectrum is directly inverted from the mel-filtered spectrum using the Moore-Penrose pseudo-inverse of the Mel-scale weighting matrix, then the inverse short-time Fourier transform magnitude (LSE–ISTFTM) algorithm is utilized to estimate the phase spectrum and recover the speech waveforms finally [10]. As for that the Mel-scale weighting matrix is wide, it demonstrates that the Mel-filter equals to an under-determined system so that there exists infinitely many solutions. The Moore-Penrose pseudo-inverse forms a least square (LS) solution, i.e., the solution has the minimum Euclidean norm. However, the physical meaning of the LS solution has not been successfully interpreted for speech processing yet. Actually, the distribution

This work is partially supported by Natural Science Foundation of China(61701535, 61871471) and Natural Science Foundation of Shanxi Province (2017JQ6033).

of the power spectrum of speech is not flat, this is because the spectral power at the harmonic frequencies and formants is apparently higher than other frequency regions. Consequently, it is reasonable to think that the power spectrum of speech is sparse. This phenomenon will provide important a prior information for recovering the power spectrum of speech much precisely. In our previous work, we used the ℓ_1 minimization technique to recover speech spectrum [11], to further exploit the sparse characteristic of speech in the frequency domain, a much simpler iteratively reweighted ℓ_2 minimization method is proposed to cope with the under-determined nature of the reconstruction problem in this paper. The quality of the reconstructed speech via this new algorithm is efficiently improved over the conventional methods used in [2]– [4].

The rest of this paper is organized as follows. In section 2, we provide a prologue that defines the problem formulation of speech reconstruction from MFCCs. Then, we propose the method of iteratively reweighted ℓ_2 minimization to recover the speech signal precisely in section 3. The results of the experimental evaluation over the TIMIT database are outlined in section 4. Finally, section 5 concludes our work.

II. Problem formulation

The extracting procedure of MFCCs begins with enframing the speech waveforms $x(n)$ by a window $w(n)$,

$$x_m(n) = x(mR+n)\,w(n) \tag{1}$$

where $L(0 \le n \le L-1)$ is the window length, R is the frame shift, m is the frame index. Then, The speech frame can be concisely denoted as,

$$\boldsymbol{x} = [x_m(0), x_m(1), ..., x_m(L-1)]^\intercal \tag{2}$$

The power spectrum of each speech frame is,

$$\boldsymbol{y} = |\mathrm{F}\{\boldsymbol{x}\}|^2 \tag{3}$$

where $\mathrm{F}\{\boldsymbol{x}\}$ is the N-point DFT of $\boldsymbol{x}$, $|\cdot|$ denotes the modulus of the complex number.

The latter $N/2-1$ elements of $\boldsymbol{y}$ will be discarded due to the symmetry. Then, the power spectrum is Mel-filtered by a set of weighting functions, i.e., the Mel-scale weighting matrix $\boldsymbol{\Phi} \in \mathbf{R}^{K\times(N/\mathbf{2}+\mathbf{1})}$, where K is the number of Mel-filter bands. Generally, $\boldsymbol{\Phi}$ is designed based on human perception of pitch frequency and implemented in the form of a bank of filters, each filter is with a triangular frequency response, as is shown in Fig. 1. Finally, MFCCs are computed through the log and discrete cosine transform,

$$\boldsymbol{f} = \mathrm{DCT}\{\log(\boldsymbol{\Phi}\boldsymbol{y})\} \tag{4}$$

The problem of speech reconstruction from MFCCs is trying to estimate $\boldsymbol{x}$ from $\boldsymbol{f}$.

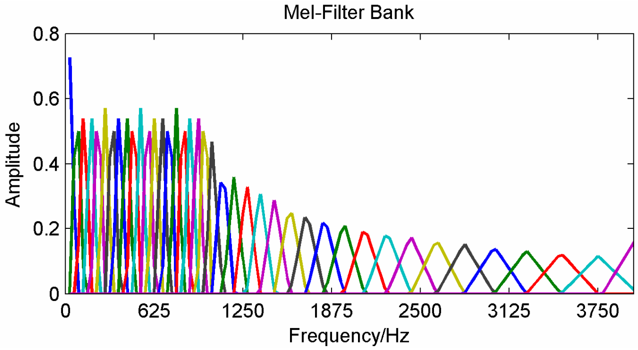

Fig. 1. The frequency response of the Mel-filter.

III. Speech reconstruction from MFCCs

A. The conventional speech reconstruction method

As is shown in (4), it is easy to see that the DCT and log operations are all invertible except for the Mel-filter. Therefore, the core of recovering the speech waveforms from MFCCs is to estimate the power spectrum. Conventionally, the ℓ_2 norm criteria is often used [2], [4],

$$(\mathrm{L}_2) \quad \min\|\boldsymbol{y}\|_2 \quad \text{subject to} \quad \boldsymbol{\Phi}\boldsymbol{y} = \boldsymbol{z} \tag{5}$$

where $\boldsymbol{z}$ is computed from $\boldsymbol{f}$ by $\boldsymbol{z} = \exp(\mathrm{IDCT}\{\boldsymbol{f}\})$, $\mathrm{IDCT}\{\cdot\}$ and $\exp(\cdot)$ denotes the inverse discrete cosine transform and element-wise exponential operation, respectively.

By solving the (L_2) problem, a minimal ℓ_2 norm solution is formed,

$$\hat{\boldsymbol{y}} = \boldsymbol{\Phi}^\dagger\boldsymbol{z} = \boldsymbol{\Phi}^\intercal(\boldsymbol{\Phi}\boldsymbol{\Phi}^\intercal)^{-1}\boldsymbol{z} \tag{6}$$

where $\boldsymbol{\Phi}^\dagger$ denotes the Moore-Penrose pseudo-inverse of $\boldsymbol{\Phi}$.

B. The proposed method

It is worth noting that the power spectrum of speech is sparse, especially for the voiced speech. The spectral power at the harmonic frequencies and formants is apparently higher than other frequency regions. This important a prior information will be beneficial for recovering the power spectrum of speech much precisely.

Inspired by the FOCUSS algorithm [12], we can recover the power spectrum exactly by minimizing the weighted ℓ_2 norm,

$$(\mathrm{WL}_2) \quad \min\|\boldsymbol{W}^{-1}\boldsymbol{y}\|_2 \quad \text{subject to} \quad \boldsymbol{\Phi}\boldsymbol{y} = \boldsymbol{z} \tag{7}$$

where $\boldsymbol{W}$ is a diagonal matrix, which aims to enhance the sparsity of the recovered power spectrum. Specifically, the diagonal elements of $\boldsymbol{W}$ are the $(1-p/2)$-power of recovered $\boldsymbol{y}$ at last iteration $(0 \le p \le 2)$.

By introducing a new variable $\boldsymbol{y}_w = \boldsymbol{W}^{-1}\boldsymbol{y}$, where $\boldsymbol{W}$ is invertible, we can rewrite (7) as follows,

$$\min\|\boldsymbol{y}_w\|_2 \quad \text{subject to} \quad \boldsymbol{\Phi}\boldsymbol{W}\boldsymbol{y}_w = \boldsymbol{z} \tag{8}$$

Then, the optimal solution of $\boldsymbol{y}_w$ is,

$$\hat{\boldsymbol{y}}_w = (\boldsymbol{\Phi}\boldsymbol{W})^\dagger\boldsymbol{z} \tag{9}$$

As for $\boldsymbol{W}$ is invertible, the closed-form solution of $\boldsymbol{y}$ in (7) is derived as,

$$\hat{\boldsymbol{y}} = \boldsymbol{W}\hat{\boldsymbol{y}}_w = \boldsymbol{W}(\boldsymbol{\Phi}\boldsymbol{W})^{\dagger}\boldsymbol{z} \tag{10}$$

In summary, the iteratively reweighted ℓ_2 minimization (IRLM) method for recovering the power spectrum is shown in **algorithm1**. To ensure the non-negativity of recovered power spectrum, we use the absolute value of $\boldsymbol{y}$ as the output.

Algorithm 1 Power spectrum reconstruction via the IRLM method.

```
Input: $\boldsymbol{\Phi}, \boldsymbol{z}$
Output: Estimate of $(\boldsymbol{y})$
1: Initialization: $\boldsymbol{W} = \mathbf{I}, \boldsymbol{y}^{(0)} = \boldsymbol{\Phi}^{\dagger}\boldsymbol{z}, k = 0, \triangle y = 1e^{8}$,
                   $p = 1, \varepsilon = 1e^{-4}, \delta = 1e^{-6}, \mathrm{M} = 20$
2: while $k \leq \mathrm{M}, \triangle y \geq \delta$ do
3:     // Line 4 updates $\boldsymbol{W}$:
4:     $\boldsymbol{W}^{(k+1)} = \mathrm{diag}\left((|\boldsymbol{y}^{(k)}| + \varepsilon)^{(1-p/2)}\right)$
5:     // Line 6 updates $\boldsymbol{y}$:
6:     $\boldsymbol{y}^{(k+1)} = \boldsymbol{W}^{(k+1)}(\boldsymbol{\Phi}\boldsymbol{W}^{(k+1)})^{\dagger}\boldsymbol{z}$
7:     // Line 8 computes the variation of $\boldsymbol{y}$, $\triangle y$:
8:     $\triangle y = \|\boldsymbol{y}^{(k+1)} - \boldsymbol{y}^{(k)}\|_2$
9:     $k = k + 1$
10: end while
11: Output $(|\boldsymbol{y}|)$
```

After the power spectrum of speech is recovered by the IRLM method, the next important step is to recover the lost phase information. Since the LSE–ISTFTM algorithm is simple, it effectively estimates the phase spectrum via modified DFT and IDFT [10]. Consequently, we also use it to recover the lost phase information. Coupling the estimated phase spectrum with the recovered amplitude spectrum will result in a time-domain estimate of the speech frame. Finally, the speech waveforms can be reconstructed via an overlap-add procedure, as is shown in Fig. 2.

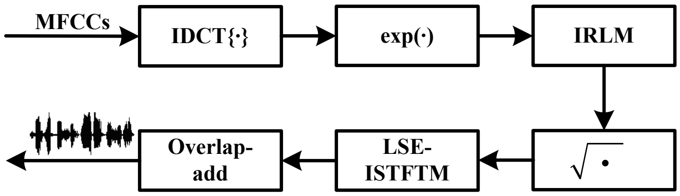

Fig. 2. *Diagram of speech reconstruction from MFCCs.*

IV. Experimental Results

A. Dataset and evaluation metrics

In the following experiments, we selected 90 speech utterances spoken by 15 males and 15 females from the TIMIT database. Each utterance is about 3 seconds in duration, which is down-sampled to 8 kHz. The speech signal is enframed to 256 samples with a hamming window, overlapped with 128 samples. Similar to [2], [4], $\boldsymbol{\Phi}$ is derived from a number of triangular weighting filters, which is linearly spaced over 0–1kHz, logarithmically spaced over 1–4kHz. We set $L = 256, R = 128, N = 256, K = 10, 20, ..., 70$, respectively. The conventional ℓ_2 minimization method (LM) used in [2]-[4] is involved for comparison in the tests.

To evaluate the proposed method objectively, four different kinds of measures are used. The average short-time objective intelligibility (STOI) score illustrates the intelligibility [13], while the perceptual evaluation of speech quality (PESQ) score illustrates the overall quality of speech [14]. The cepstral distance and frequency-weighted segmental SNR (fwsegSNRs) are two other popular objective measures for speech processing [15]. Except for the cepstral distance measure, higher value indicates better performance. To encourage reproducing similar experimental results, we provide the source codes here: $https : //ww2.mathworks.cn/matlabcentral/fileexchange/$ $53186 - invmfccs$.

B. Parameter specification

In algorithm1, there are four parameters to specify, i.e. p, ε, δ and M. Empirically, the maximum number of iterations M and the error tolerance δ are specified as 20 and $1e^{-6}$, respectively, which is adequate for the IRLM method to achieve a stationary solution. In addition, ε should be specified so that the weighting matrix $\boldsymbol{W}$ avoids being singular. We specify it as $1e^{-4}$ empirically. Given different choices of p, we conduct the experiments over the dataset described in 4.1. The experimental results are shown in Fig. 3.

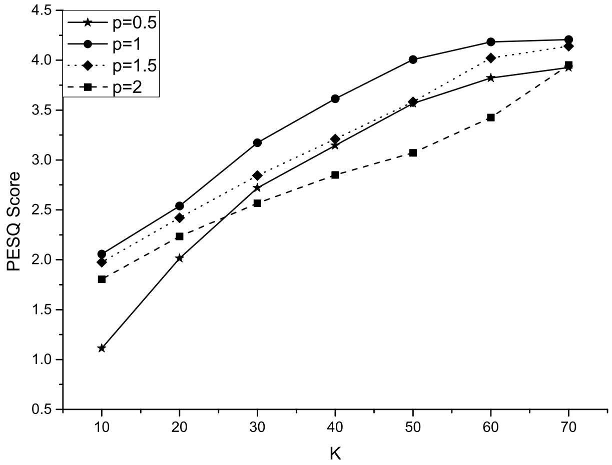

Fig. 3. *Comparison on the quality of the recovered speech in terms of PESQ score for different choice of p.*

From Fig. 3, we can see that $p = 1$ gives the best speech quality in terms of PESQ score. Therefore, p will be fixed to this value for the IRLM method in the following experiments.

C. Performance evaluation

The results of the performance evaluation using the objective measures are shown in Figures 4–5 and Tables 1–2. It is illustrated that the IRLM method achieves substantially higher fwsegSNRs, STOI, PESQ score and lower cepstral distance than the conventional LM method, which demonstrates that the speech quality is much better. Especially, the improvement is dramatic for female utterances. For instance, the average fwsegSNRs improvement is 0.803dB for male utterances while it is 2.778dB for female utterances, the average PESQ score improvement is 0.31 for male utterances while it is 0.94 for female utterances. The main reason for the gender differences is that the female utterance is much sparser in the frequency domain. However, the conventional LM method does not fully exploit the sparse characteristic through optimizing the ℓ_2 norm. When observing the spectrogram of the speech recovered by the LM method, we will find that the harmonic structure, especially in the low frequency region, is severely "smeared". The smeared amplitude spectrum degrades the articulation of the reconstructed speech.

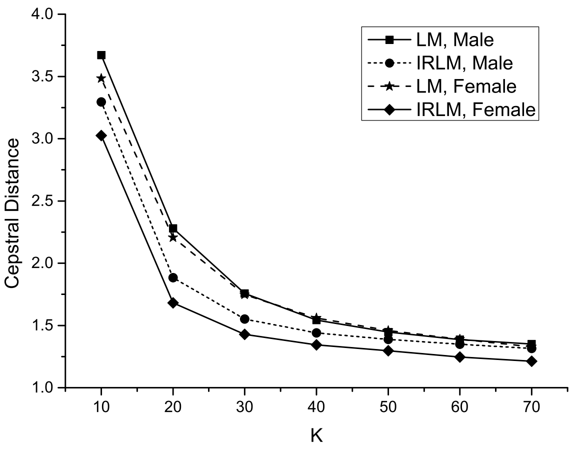

Fig. 4. *Comparison on the cepstral distance.*

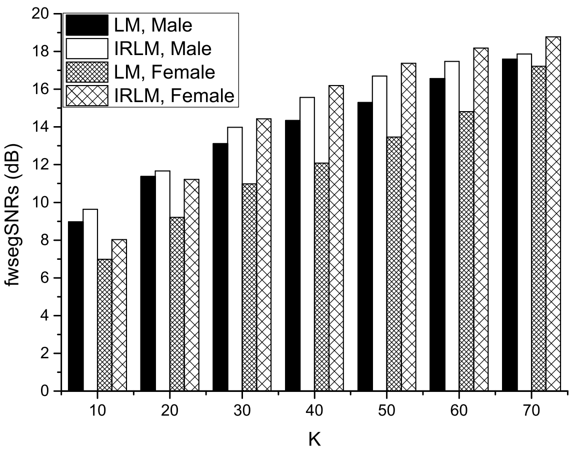

Fig. 5. *Comparison on the fwsegSNRs.*

TABLE I
Comparison on the STOI score (%).

K	LM			IRLM		
	Male	Female	Avg.	Male	Female	Avg.
10	76.40	72.29	74.35	79.89	76.53	78.21
20	83.00	79.28	81.14	84.90	86.91	85.91
30	86.18	83.18	84.68	88.54	93.22	90.88
40	88.89	85.95	87.42	91.28	94.93	93.11
50	90.08	88.18	89.13	92.55	95.41	93.98
60	91.53	90.17	90.85	93.23	95.93	94.58
70	93.25	95.63	94.44	93.44	96.13	94.79

TABLE II
Comparison on the PESQ score.

K	LM			IRLM		
	Male	Female	Avg.	Male	Female	Avg.
10	2.197	1.431	1.805	2.510	1.602	2.056
20	2.645	1.821	2.233	2.798	2.277	2.538
30	2.953	2.180	2.567	3.216	3.129	3.173
40	3.275	2.422	2.849	3.720	3.506	3.613
50	3.483	2.658	3.071	4.061	3.952	4.007
60	3.851	2.998	3.425	4.234	4.134	4.184
70	4.222	3.682	3.952	4.270	4.144	4.207

In addition, we also conduct the subjective A/B listening tests by four listeners. The test results also demonstrate that speech recovered by the IRLM method is clearer and intelligible than those corresponding to the LM mehtod. Also, the recovered speech sounds very close to the original speech when using the high-resolution MFCCs. For instance, when $K = 50$, the STOI, PESQ score is beyond 93% and 4.0, respectively. Consequently, it is reasonable to believe that the sparse priors is very important for speech reconstruction. In reality, the IRLM method successfully exploits the sparse characteristic through iteratively reweighted optimization. As a result, it achieves substantially better performance.

V. Conclusion

In this paper, we propose a simple and effective method to recover speech from MFCCs. This method successfully exploits the sparse characteristic of the speech spectrum via iteratively reweighted ℓ_2 minimization. Extensive evaluations over the TIMIT database have shown that the quality of the recovered speech is efficiently improved when compared to the output of the conventional method. Experimental results also verify that the sparse priors of speech in the frequency domain is important for speech reconstruction with high quality. This proposed method is easily embed into MFCC codec.

It is expected that the performance could be improved further. For example, speech reconstruction from MFCCs or mel-spectrum using neural networks is currently an active topic in speech synthesis and low rate coding [16] [17]. Also, the LSE–ISTFTM speech waveform synthesizing method is ripe for improvement since the LSE–ISTFTM outputs may suffer from audible artifacts. Consequently, we are planning to explore high-quality mel-spectrum inversion method using deep neural networks in future work.

Acknowledgment

The authors would like to thank Prof. Phillip L. De Leon and Dr. Steven Sandoval for constructive discussions and permission of using MFCC codec source code. The authors would also like to thank the anonymous reviewers for valuable suggestions which helped to improve the quality of this paper.

REFERENCES

[1] T. F. Quatieri, Discrete Time Speech Signal Processing, Upper Saddle River, NJ:Prentice-Hall, 2002.

[2] L. E. Boucheron, P.L. De Leon, and S. Sandoval, “Low Bit-Rate Speech Coding Through Quantization of Mel-Frequency Cepstral Coefficients,” IEEE Trans. Audio, Speech, and Language Processing. vol. 20, no. 2, pp. 610-619, Feb. 2012.

[3] G. Min, X. W. Zhang, X. Zou, etc “Percetually weighted analysis-by-synthesis vector quantization for MFCC Codec,” IEEE Signal Processing Letters. vol. 23, no. 10, pp. 1379-1383, Oct. 2016.

[4] L. E. Boucheron and P.L. De Leon, “On the inversion of Mel-frequency cepstral coefficients for speech enhancement applications,” in Proc. ICSES. IEEE, 2008, pp. 485-488.

[5] D. Chazan, R. Hoory, G. Cohen, etc, “Speech reconstruction from mel frequency cepstral coefficients and pitch frequency,” in Proc. ICASSP. IEEE, 2000, vol. III, pp. 1299-1302.

[6] X. Shao and B. Milner, “Clean speech reconstruction from noisy mel-frequency cepstral coefficients using a sinusoidal model,” in Proc. ICASSP. IEEE, 2003, vol. I, pp. 704-707.

[7] ETSI ES 202 211, “Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Extended front-end feature extraction algorithm; Compression algorithms; Back-end speech reconstruction algorithm,” 2003.

[8] ETSI ES 202 212, “Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Extended advanced front-end feature extraction algorithm; Compression algorithms; Back-end speech reconstruction algorithm,” 2005.

[9] B. Milner and X. Shao, “Prediction of fundamental frequency and voicing from mel-frequency cepstral coefficients for unconstrained speech reconstruction,” IEEE Trans. Audio, Speech, and Language Processing. vol. 15, no. 1, pp. 24-33, Jan. 2007.

[10] D. W. Griffin and J.S. Lim, “Signal estimation from modified short time fourier transform,” IEEE Trans. Audio, Speech, Language Processing. vol. 32, no. 2, pp. 236-243, April. 1984.

[11] G. Min, X. W. Zhang, J. B. Yang, etc, ”Speech reconstruction from mel-frequency cepstral coefficients via ℓ_1-norm minimization,” in Proc. MMSP. IEEE, 2015, pp. 1-5.

[12] I. F. Gorodnitsky and B.D. Rao, “Sparse signal reconstruction from limited data using FOCUSS: A reweighted minimum norm algorithm,” IEEE Trans. Signal Processing. vol. 45, no. 3, pp. 600-616, Mar. 1997.

[13] C. H. Taal, R. C. Hendriks, R. Heusdens, etc, “An Algorithm for intelligibility prediction of time-frequency weighted noisy speech,” IEEE Trans. Audio, Speech, and Language Processing. vol. 19, no. 7. pp. 2125-2136, Jul. 2011.

[14] A. W. Rix, J. G. Beerends, M. P. Hollier, etc, “Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs,” in Proc. ICASSP. IEEE, 2001, vol. II, pp. 749-752.

[15] J. Tribolet, P. Noll, B. McDermott, etc, ”A study of complexity and quality of speech waveform coders,” in Proc. ICASSP. IEEE, 1978, pp. 586-590.

[16] W. B. Kleijn, F. S. C. Lim, A. Luebs, etc, ”Wavenet Based Low Rate Speech Coding,” in Proc. ICASSP. IEEE, 2015, pp. 676-680.

[17] L. Juvela, B. Bollepalli, X. Wang, etc, ”Speech waveform synthesis from MFCC sequences with generative adversarial networks,” in Proc. ICASSP. IEEE, 2018, pp. 5679-5683.