Speech Enhancement for Demodulated Signals under Multipath Fading Communication Channels

Akio Kobayashi*
* University of Tsukuba Technology, Tsukuba, Japan
E-mail:a-kobayashi@a.tsukuba-tech.ac.jp

***Abstract*—In analog communication channels, such as radio broadcasting, the superposition of multiple reflected signals causes multipath fading. Multipath fading often results in the fluctuation of the received electric intensity levels of these signals; thus, it causes severe quality degradation in audible sounds. In this paper, we focus on speech enhancement under a fading communication channel with additive Gaussian noise. We attempt to reconstruct the original speech based on the use of denoising autoencoders that employ mean-squared-error and additive perceptual evaluation of speech quality (PESQ)-based loss functions in multi-task learning (MTL). The experimental results indicate that the MTL-based autoencoder improves PESQ scores from 2.00 to 2.75 for demodulated signals under fading communication channels with additive Gaussian noise.**

I. Introduction

Although analog radio broadcasting based on wireless communication technology is generally considered a legacy media, it continues to occupy a prominent position in daily life. Specifically, in Japan, it is recognized as a means of receiving information in disasters, such as floods and mega-earthquakes [1]. There are two reasons why such obsolete media is still preferred. First, analog radio receivers are available at a lower cost than digital devices, such as smartphones. Second, the low battery consumption of analog receivers enables the long service periods required during extended blackout situations. In radio broadcasting, a variety of receiving environments exist, such as urban areas surrounded by high-rises and suburban regions where open fields are typical. The quality of the received signals principally depends on the electric intensity levels at the receiving points; they considerably fluctuate owing to the interference of reflected waves and the existence of additive Gaussian noise in the communication channel. In a wireless communication channel, one can observe the degradation of receiving signals based on low signal-to-noise ratios (SNRs) and fluctuations caused by reflections occurring in the ionosphere or obstacles, such as mountains. The fluctuations caused by these types of reflectors are called *multipath fading* [2]. Multipath fading is a phenomenon generated by the superposition of multiple reflected signals having different amplitudes and phases. In this case, the electric intensity levels vary drastically depending on the arrival time differences of the reflected signals at the receiving point. Therefore, if we introduce speech enhancement approaches to this form of media, it is possible to automatically measure the signal quality over a wide area of receiving points and place a relay station at a location where it can recover the degraded signals. Since it is unnecessary to evaluate the quality of signals manually at the receiving point, it will enable us to place a relay station at an overwhelmingly low cost. In the research field of speech enhancement, there have been many attempts to reconstruct the original signal in a noisy environment. In the literature, for example, denoising autoencoders [3], [4], [5], [6], [7], variational-autoencoders [8], and generative-adversarial-network-based methods [9], [10] have been proposed, and these approaches have been used to obtain high-quality signals from degraded ones. However, these attempts have only been applied to signals in noisy environments. There have not been attempts to reconstruct fluctuating signals that are subject to multipath fading. In speech enhancement, perceptual evaluation of speech quality (PESQ) [11] generally measures the quality of the reference/degraded signals and estimates the mean opinion score (MOS) by comparing them; thus, there is a drawback in that it is not practically portable because estimating the MOS requires two time-aligned signals. In recent years, non-intrusive PESQ estimation methods based on neural networks have been proposed to counter this drawback [12].

In this paper, we attempt to recover speech signals affected by multipath fading based on the use of denoising autoencoders. We propose a method based on multi-task learning, which incorporates mean squared errors (MSEs) and PESQ scores. Then, we discuss the efficacy of autoencoder-based speech enhancement under fading channel conditions.

II. Communication Channels in Radio Broadcasting

In this section, we briefly describe wireless communication in radio broadcasting, followed by amplitude modulation (AM), one of the popular global signal modulation methods. Then, we outline a fading communication channel, which degrades the quality of signals.

A. Wireless Communication Based on Amplitude Modulation

In analog radio broadcasting, AM and frequency modulation (FM) are the mainstream modulation methods used worldwide, and this study focuses on AM, which is easy to implement and simulate using software-defined radio tools.

Fig. 1 shows the modulation/demodulation diagram for the wireless communication channel discussed in this paper. The modulated signals pass through the fading communication channel with additive Gaussian noise (AGN). At the receiving

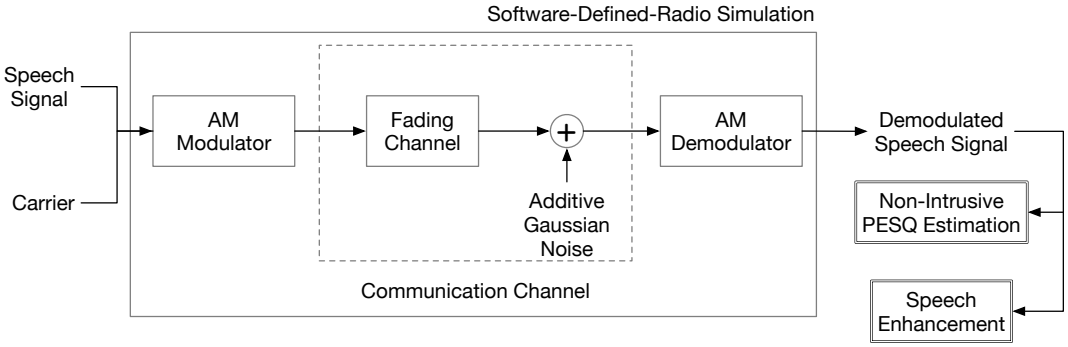

Fig. 1. Overview of AM Wireless Communication Channel Simulation and Speech Enhancement

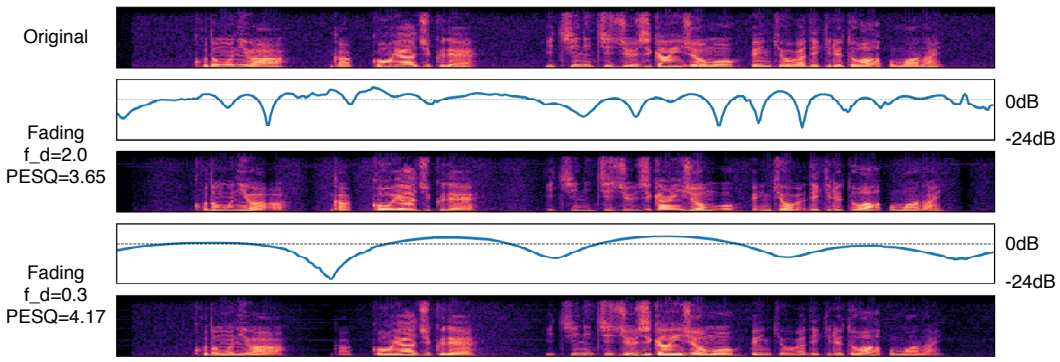

Fig. 2. Examples of Demodulated Speech Signals under Multipath Fading

point, the signals are demodulated. For simplicity, we do not consider any propagation losses through the communication channel. The AM carrier frequency ranges from 300 kHz to 3 MHz, and this frequency band is called the medium frequency (MF) [13].

In AM, the carrier without a phase offset is defined as

$$v_c(t) = V_c \cos 2\pi f_c t, \tag{1}$$

where V_c is the carrier voltage and f_c is the carrier frequency.

For simplicity, consider a speech signal represented by a cosine wave

$$v_s(t) = V_s \cos 2\pi f_s t, \tag{2}$$

where V_s is the signal voltage and f_s is the signal frequency.

The modulation is

$$\begin{aligned} v_m(t) &= (V_c + V_s \cos 2\pi f_s t) \cos 2\pi f_c t \\ &= V_c \cos 2\pi f_c t + \frac{V_s}{2} (\cos 2\pi (f_c + f_s) t \\ &\quad + \cos 2\pi (f_c - f_s) t) . \end{aligned} \tag{3}$$

As indicated by the equation, one of the characteristics of AM is that modulation signals appear on the upper and lower sides of the carrier frequency in the spectrum.

In this study, we simulate the AM communication channel according to GNU Radio [14], which is a software-defined radio toolkit. In software-defined radio, demodulation is realized by a simple cascade of decimation, band limitation, and low-pass filters.

B. Fading Communication Channels

In wireless communication, the signal emitted from the transmission point can reach the receiving point through multiple paths. That is, buildings, mountains, and the ionosphere can act as reflectors, and therefore multiple signals can be observed at the receiving point. At this time, the signal is attenuated by amplitude/phase differences among the received signals. This phenomenon is called multipath fading [2], [15]. When the reflected signals overlap with each other, the composite signal is approximated by the Rayleigh distribution:

$$f(x) = \frac{x}{\sigma^2} \exp\left(-\frac{x}{2\sigma^2}\right), \tag{4}$$

where $\sigma^2 = E[x^2]$. Multipath fading often causes drastic changes in the amplitudes of demodulated signals, which can significantly degrade the sound quality. In addition, in ionospheric reflection, the reflected signals may experience Doppler frequency shifts because of the fast movement of molecules in the ionosphere [16].

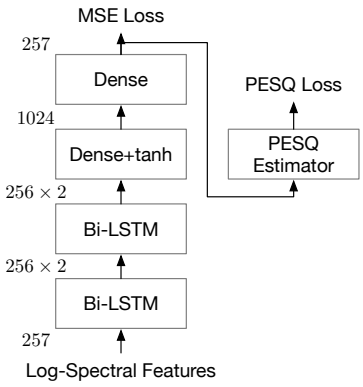

Fig. 3. Denoising Autoencoder Configuration for Speech Enhancement

In this study, we assume that the signal in the communication channel is affected by a Doppler frequency shift caused by the motion of molecules. Then, we simulate the signal in the fading channel according to [17]. The complex gains for the modulated signal, which follows the Rayleigh distribution, with the maximum Doppler shift f_d, are given by

$$x(t) = \sqrt{\frac{2}{N}} \sum_{n}^{N} \exp\left(j\left(2\pi f_d t \cos\alpha_n + \phi_n\right)\right), \qquad (5)$$

where N is the number of sinusoids and α_n, and ϕ_n are phases.

Fig. 2 provides examples of demodulated signals under a fading channel. From top to bottom, the magnitude ratios of the received/original signals without propagation loss, and the spectrograms of the original signals and those for the received signals, are shown in the cases of $f_d = 2.0$ and $f_d = 0.3$. In Fig. 2, we can observe sharp drops in the signal levels. If such signal drops are present, the quality during listening degrades significantly.

C. PESQ

While the level crossing rate (LCR) and average fading duration (AFD) are typical measurements used to evaluate fading [18], PESQ [11], STOI [19], and log-spectral distortion(LSD) [20] are broadly used as objective quality measures in the research field of spoken language processing [21]. Therefore, we employ PESQ as a measure of speech quality of demodulated signals under fading communication channels.

PESQ is a standard quality assessment tool for speech enhancement, but it has the drawback of lacking portability because it requires two signal sources that can be aligned in the temporal direction. In recent years, PESQ estimation using neural networks has been studied to overcome this drawback [22]. By contrast, an alternate method that employs PESQ as an additive loss function has been proposed [23]. In this study, we estimate the PESQ without reference using a neural network and discuss the efficacy of PESQ estimation for degraded speech under fading channel conditions.

TABLE I
PESQ ESTIMATION RESULTS

	MSE	corr.
w/o AGN	0.140	0.83
w/ AGN	0.076	0.98

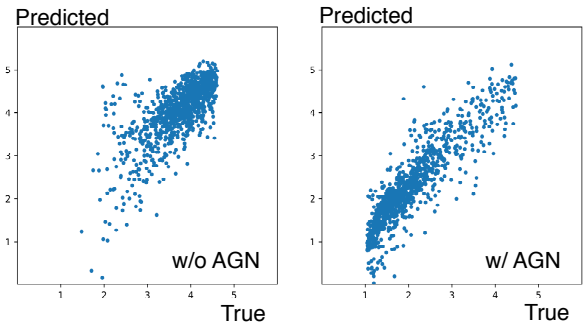

Fig. 4. Scatter Plots of True/Predicted PESQ Scores

D. Denoising Autoencoder

Our question is whether the original signal can be reconstructed from the degraded signals in a fading communication channel. A denoising autoencoder is generally used to restore the degraded signal and improve its quality. To date, many approaches have been proposed [3], [4], [5], [6], [7]. Many of these approaches have attempted to recover the original signals from those under noisy environments. However, there are few studies focused on speech signals whose levels fluctuate significantly. Therefore, we first attempt to recover speech in the fading channel using denoising autoencoders based on bidirectional long short-term memory (LSTM) layers. In general, the denoising autoencoder minimizes the mean squared error between the target signal and the predicted signal; however, in this study, we attempt multi-task learning (MTL) by incorporating PESQ into the objective function, similar to [23].

III. EXPERIMENTAL SETUP

In this section, we briefly summarize the experimental setups.

A. Corpus

In the experiments, we used JNAS [24], a Japanese speech corpus, consisting of over 40 k utterances at a 16-kHz sampling rate. We divided the corpus into two subsets for training and validation so that there was no overlap between speakers and utterances. We prepared a total of five pairs of training/validation sets; the training sets consisted of 22 k utterances on average, and the validation sets consisted of 200 utterances. In the following experiments, we conducted five-fold cross-validation to obtain the experimental results.

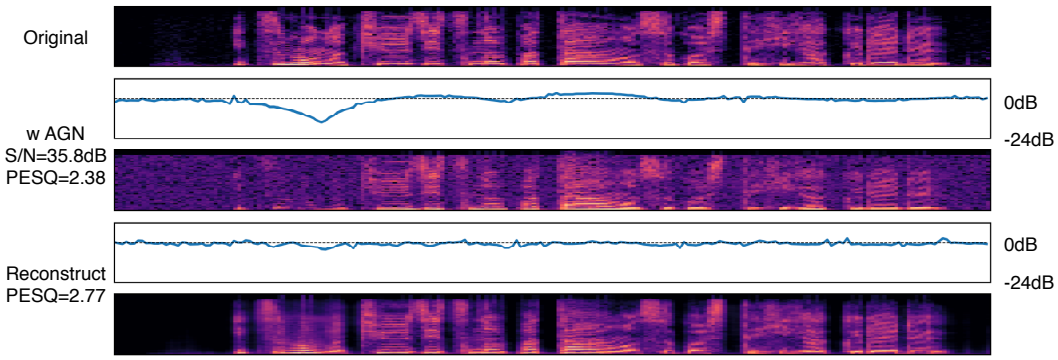

Fig. 5. Examples of Reconstructed Speech Signals

B. Pseudo Data in Fading Communication Channel

We prepared a set of pseudo signals in the fading communication channel with additive Gaussian noise (AGN) according to the GNU Radio [14],[17] channel simulator, or one of the software-defined radio tools. When configuring the fading channel, we set the maximum Doppler frequency, which ranged randomly from 0.0 Hz to 2.0 Hz at a carrier frequency of 594 kHz, and simulated eight sinusoids utilized in Eq.(5) to generate the pseudo signals. In this simulated channel configuration, the PESQ scores of pseudo signals ranged from 1.30 to 4.64 and averaged 3.78. Moreover, in the simulated channel with AGN, the PESQ scores ranged from 1.03 to 4.59 and averaged 2.53, while the SNRs ranged from 6.75 dB to 56.0 dB. The PESQ scores of the original signals were set to 4.5.

C. Network Configurations

In the non-intrusive PESQ estimation, we employed the Quality-Net (QN) proposed in [22]. From the results of the preliminary experiments, we utilized 257-dimensional logarithmic spectral features as inputs, while the original QN employed non-logarithmic ones. The network, as well as the original QN, is configured as a cascade of a bidirectional LSTM with 100-dimensional cell units, a dense layer followed by an exponential linear unit layer, and a global average pooling layer for taking PESQ scores. In the original QN training scheme, MTL consists of two loss functions: the frame-averaged PESQ loss over the entire frame, and the frame-wise loss. MTL was performed with fixed loss weights. In this study, we set the weight of the frame-wise loss to 0.1.

In the speech enhancement experiment, we used the denoising autoencoder configuration shown in Fig. 3. To output logarithmic spectral features, the network combines two bidirectional LSTM layers: a fully connected layer followed by hyperbolic tangent activations and a fully connected layer (**Baseline**). Meanwhile, we connect the QN for PESQ estimation to the output layer as an additional loss function, and training is performed by multi-task learning, which is referred to as **MTL**. The PESQ estimator parameters are fixed, and the additional loss function is minimized so that the estimated PESQ scores are converged to 4.5.

TABLE II
EVALUATION RESULTS (W/O AGN)

	MSE	PESQ	STOI	LSD
Demodulated	n/a	3.22	0.90	17.9
Baseline	0.0113	3.36	0.92	15.7
MTL	0.0107	3.38	0.92	15.6

TABLE III
EVALUATION RESULTS (W/ AGN)

	MSE	PESQ	STOI	LSD
Demodulated	n/a	2.00	0.87	18.4
Baseline	0.0485	2.74	0.87	16.8
MTL	0.0487	2.75	0.87	16.8

IV. EXPERIMENTAL RESULTS

A. PESQ Estimation

First, we describe the results of the PESQ estimation in fading communication channels. For a fading channel and channel with AGN, Table I lists the MSEs and the correlations between true and predicted PESQ scores. As shown in Table I, the model for the fading channel without AGN yielded a larger MSE than that for the channel with AGN. As shown in Fig. 2, multipath fading tends to significantly degrade the signal in a local narrow range. Therefore, it would be difficult for the QN to estimate the PESQ from such a local decline. By contrast, because most real-world signals exist with additive Gaussian noise, PESQ estimation does not matter in reality.

Fig. 4 shows scatter plots of the true PESQ against the predicted one. These plots demonstrate that the QN (the non-intrusive model), can estimate the PESQ for degraded speech signals under fading channels. On the other hand, when the true PESQ scores are low, the predicted scores tend to be underestimated.

B. Denoising Autoencoder Results

Tables II & III show measured results of demodulated signals in both channels according to PESQ, STOI, and LSD. All results are the average results obtained from five-fold cross-validation sets, and the predicted logarithmic spectral features were converted into time-domain signals using the Griffin–Lim algorithm [25], [26].

In Table II, the **Baseline** autoencoder improved the average PESQ score from 3.22 to 3.36 compared to the **Demodulated** approach, while **MTL** achieved a PESQ score of 3.38, which was a 0.6% improvement compared to the **Baseline**. PESQ scores decreased for signals in the communication channel with AGN. **MTL** achieved a slightly better score compared to the **Baseline** and improved the PESQ by 0.4% compared to the demodulated signals. **MTL** achieved small improvements in PESQ scores against the **Baseline**. This is probably because the decline of the signal levels due to fading occurs only in a small portion of the signals.

Although in Table III, there are no improvements in STOI, this type of measure is not affected by a small portion of extreme signal distortion, as seen in fading, since STOI is calculated from the average of local distortions[19].

Fig. 5 shows a sample of the demodulated signal and the reconstructed signal after applying the denoising autoencoder. The reconstructed signal was adjusted in gain to the original gain and then recovered from the dip caused by a short temporal span in level-drop.

V. Conclusion

In this paper, we investigated a speech enhancement method for reconstructing demodulated speech signals under a multipath fading channel typically found in radio broadcasting. We simulated two communication channels, one with fading only and the other with additive Gaussian noise and generated pseudo data sets. Experimental results indicate that the neural networks based on bi-directional LSTMs improved the PESQ value in the fading channel with Gaussian noise from 2.00 to 2.75. Encouraged by these promising results, we would like to design a corpus in future work that includes speech signals under fading channels in the real world and evaluate the results based on subjective assessment methods. In a practical situation, the subjective assessment is employed as a measure of quality. Thus, it is necessary to aim for a system using the subjective assessment instead of PESQ.

Acknowledgement

This work was supported by JSPS KAKENHI, Grant Number JP19096334.

References

[1] Ministry of Internal Affairs and Communications "Information and communications in Japan (White Paper),", 2019, https://www.soumu.go.jp/johotsusintokei/whitepaper/eng/WP2019/2019-index.html

[2] H. Bai and M. Atiquzzaman, "Error modeling schemes for fading channels in wireless communications: A survey," in *IEEE Communications Surveys Tutorials*, 2003, vol.5, no.2, pp. 2–9.

[3] X. Liu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in Proc. *Interspeech*, 2013,pp.436–439.

[4] P.G. Shivakumar, P. Georgiou, "Perception optimized deep denoising autoencoders for speech enhancement," in Proc. *Interspeech*, 2016, pp.3743–3747.

[5] F.-K. Chuang, S.-S. Wang, J.-W. Hung, Y. Tsao, and S.-H. Fang, "Speaker-aware deep denoising autoencoder with embedded speaker identity for speech enhancement," in Proc. Inetrspeech, 2019, pp.3173–3177.

[6] N. Tawara, T. Kobayashi, and T. Ogawa, "Multi-channel speech enhancement using time-domain convolutional denoising autoencoder," in Proc. *Interspeech*, 2019, pp.86–90.

[7] M. Liu, Y. Wang, J. Wang, J. Wang, and X. Xie, "Speech enhancement method based on LSTM neural network for speech recognition," in Proc. 14th IEEE International Conference on Signal Processing, 2018, pp. 245–249.

[8] Y. Bando, M. Mimura, K. Itoyama, K Yoshii, and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," in *Proc. ICASSP*, 2018, pp.716–720.

[9] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," *Proc. Interspeech 2017*, 2017, pp.3642–3646.

[10] M. H. Soni,N Shah, and H. A. Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," in *Proc. ICASSP*, 2018, pp.5039–5043

[11] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ): A new method for speech quality assessment of telephone networks and codecs," in *Proc.ICASSP 2001*, 2001, pp. 749—752.

[12] A. R. Avila, H. Gamper, C. Reddy R. Cutler I. Tashev, and J. Gehrke, "Non-intrusive speech quality assessment using neural networks," in *Proc. ICASSP 2019*, 2019, pp.631–635.

[13] M. Sibley, "Modern telecommunications: Basic principles and practices,", CRC Press, 2018.

[14] T. J. O'Shea, and N. E. West, "Radio machine learning dataset generation with GNU radio, in *Proc. GNU Radio Conference*, 2016.

[15] P. Barsocchi, "Channel models for terrestrial wireless communications: a survey," in *CNR-ISTI technical report*, 2006.

[16] C. Bianchi, J. A. Baskaradas, M. Pezzopane, M. Pietrella, U. Sciacca, and E. Zuccheretti, "Fading in the HF ionospheric channel and the role of irregularities," In *Advances in Space Research*, 2013, vol.52, pp.403–411.

[17] A. Alimohammad, S. F. Fard, B. F. Cockburn, and C. Schlegel, "Compact Rayleigh and Rician fading simulator based on random walk processes," in *IET Communications*, 2009, vol.3, 1333–1342.

[18] Z. Hadzi-Velkov, N. Zlatanov, and G. K. Karagiannidis, "Level crossing rate and average fade duration of the multihop Rayleigh fading channel," in *Proc. IEEE International Conference on Communications*, 2008, 4451–4455.

[19] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. ICASSP*, 2010, pp. 4214—4217.

[20] A. Prodeus and I. Kotvytskyi, "On reliability of log-spectral distortion measure in speech quality estimation," in *Proc. APUAVD*,2017, pp. 121–124.

[21] "Evaluation of objective quality measures for speech enhancement," in *IEEE Trans. ASSP*, 2008, vol.16, no.1, pp.229-–238.

[22] S.-W. Fu, Y. Tsao, H.-T. Hwang, and H. M. Wang, "Quality-net: An end-to-end non-intrusive speech quality assessment model based on BLSTM," in *Proc. Interspeech*, 2018, pp. 1873–1877.

[23] S. Fu, C. Liao, and Y. Tsao, "Learning with learned loss function: Speech enhancement with quality-net to improve perceptual evaluation of speech quality," in *IEEE Signal Processing Letters*, 2020, vol.27, pp. 26–30.

[24] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," in *J. of Acoustic Society of Japan (E)*, 1999, vol.29, no.3, pp. 199-206.

[25] D. W. Griffin, and J. S. Lim, "Signal estimation from modified short-time Fourier transform," in *IEEE Trans. ASSP*, 1984, vol.32, no.2, pp.236-–243.

[26] N. Perraudin, P. Balazs, and P. L. Søndergaard, "A fast Griffin-Lim algorithm," in Proc. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013, pp. 1–4.