A Study on Sampling of STFT Modifications in Time and Frequency Domains for DNN-Based Speech Dereverberation

Bo Wu*and Kehuang Li† and Minglei Yang* and Chin-Hui Lee†
* National Laboratory of Radar Signal Processing, Xidian University, Xi'an, China
E-mail: rambowu11@gmail.com; mlyang@xidian.edu.cn
† School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, USA
E-mail: kehle@gatech.edu; chl@ece.gatech.edu

Abstract—**We investigate the effects of time and frequency sampling on short-time Fourier transform modifications to be used for speech dereverberation based on deep neural networks (DNNs). We first show that by adopting a linear activation function at the output layer and globally normalizing the target features into zero mean and unit variance, better performances can be obtained than existing DNN approaches. Then we show that the quality of dereverberated speech could be degraded with denser sampling in time for longer reverberation times, even at the price of increased computational complexities, requiring an adaptive time sampling strategy. On the other hand, the difference between the unwrapped phases of reverberant and anechoic speech becomes negligible with a dense sampling in frequency, implying a reduced speech distortion. Therefore, there is a great potential to enhance DNN based acoustic signal processing if the conventional sampling strategy can be carefully adjusted.**

I. Introduction

When a microphone is placed at a distance from a talker in an enclosed space, the received signal will be a collection of many delayed and attenuated copies of the original speech signals, that are caused by the reflections from walls, ceilings, or floors [1]. As a result, reverberation often degrades speech quality and intelligibility.

Many dereverberation techniques have been proposed in the past [2], [3], [4], [5]. One direct way is to estimate an inverse filter of the room impulse response (RIR) [6] to deconvolve the reverberated signal [2]. However, a minimum phase assumption is often needed, which is almost never satisfied in practice [6]. The RIR can also be varying in time and hard to estimate [1]. Recently, due to their strong regression capabilities, deep neural networks (DNNs) have been widely used in speech enhancement [7], source separation [8], and bandwidth expansion [9]. Han *et al.* [5] also proposed to dereverberate speech using DNNs, to learn a spectral mapping from reverberant to anechoic speech. Although good results have been reported, they utilized a sigmoid activation function at the output layer and normalized the target feature into an unit range, restricting the performance improvements. While we proposed to adopt a linear activation function at the output layer and to globally normalize the target features into zero mean and unit variance in [10], achieving considerable performance improvements, especially for perceptual evaluation of speech quality (PESQ) [11] .

Most dereverberation algorithms use the short-time Fourier transform (STFT) [12], which is sampled in both time and frequency dimensions, to obtain a discrete time-frequency representation of speech. The conventional sampling strategy is that the rates are chosen to avoid aliasing in both time and frequency domains, in order to reconstruct exactly the speech signal from its corresponding sampled STFT [13]. Since the DNN-based approach is also based on an analysis-modification-synthesis (AMS) framework, the sampling strategy should take into consideration not only reconstruction, but also the modification procedures.

In this study, we investigate the impact of time and frequency sampling rates of STFT on the DNN-based dereverberation performances. A recently proposed DNN model [10] is utilized to learn a high-quality performance. We show that for time domain sampling, denser sampling can result in a degradation of DNN-based dereverberated speech for longer reverberation times, even though speech can be reconstructed more exactly. For frequency domain sampling, the estimation of the unwrapped phase depends on the frequency sampling rate. With a denser sampling, the difference between unwrapped phases of reverberant and anechoic speech becomes less significant.

II. DNN-based Speech Dereverberation

A block diagram of the DNN-based speech dereverberation system, proposed in [10], is illustrated in Fig. 1. In the training stage, the DNN, as a regression model, is trained by using log-power spectral (LPS) features from pairs of the reverberant and anechoic signals from a few input frames. In the dereverberation stage, the well-trained DNN model is fed with the LPS features of unseen speech to estimate the LPS features of anechoic speech. Then we utilize the estimated spectral magnitude and reverberant speech's phase to reconstruct the estimated anechoic waveform with an overlap-add method.

In feature extraction, the spacing between adjacent analysis window positions is the frame shift R and the N-point discrete

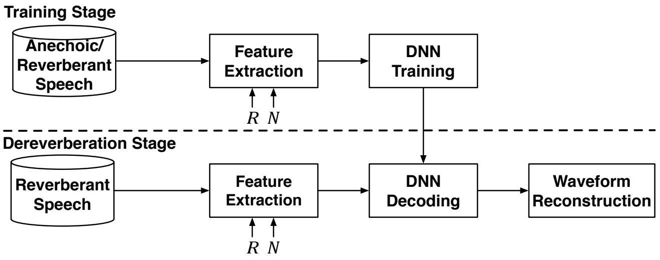

Fig. 1. A DNN-based speech dereverberation system.

Fourier transform (DFT) is computed for each overlapping windowed frame. Note that, the frame shift R is called the time sampling rate of STFT in [13], different from the time sampling rate T_s of the speech signal, $x(n)$. And N denotes the frequency sampling rate.

III. Sampling and Modification of STFT

The STFT is a function of both the discrete time n and continuous normalized radian frequency ω, defined as [13]:

$$X_n(e^{j\omega}) = \sum_{m=-\infty}^{\infty} w(n-m)x(m)e^{-j\omega m} \tag{1}$$

where $w(n-m)$ is a real time-shifted window to get a portion of the input signal $x(n)$ at a particular time index n with a window of length L.

A. Time Sampling of STFT

Eq. (2) samples STFT at a time rate of (i.e., frame shift) R,

$$X_{rR}(e^{j\omega}) = \sum_{m=-\infty}^{\infty} w(rR-m)x(m)e^{-j\omega m} \tag{2}$$

where r denotes frame index.

In conventional time sampling of STFT, the frame shift is chosen to avoid an aliased representation of $X_n(e^{j\omega})$ from which $x(n)$ can be exactly recovered [14]. The frame shift is typically fixed to half of the frame length [15] for practical consideration in most dereverberation algorithms.

However, in a reverberant environment, the conventional method is ineffective [2], [3], [4], [5] because it uses a fixed temporal resolution, without considering the mixing conditions of neighboring reverberant frames at different RT60s. For weak reverberation, the reflected sounds travel a less distance to a microphone [16], resulting in an intensive superposition in the time domain. A dense sampling of $X_n(e^{j\omega})$, not needed for strong reverberation, is now required to provide a high temporal resolution.

B. Frequency Sampling of STFT

STFT, sampled at frequencies $\Omega_k = 2\pi k/N$ $(k = 0, 1, ..., N-1)$ as the N-point DFT of the time-limited sequence $w(n-m)x(m)$, with N denoting the sampling rate:

$$X_n(e^{j\Omega_k}) = \sum_{m=-\infty}^{\infty} w(n-m)x(m)e^{-j\Omega_k m}. \tag{3}$$

The conventional frequency sampling strategy is to avoid aliasing. According to the sampling theorem [17], $X_n(e^{j\omega})$ should be sampled at a rate of at least twice its "time-width", i.e., $N \geq L$ [13]. In order to reduce the computational complexity, most of the DNN-based systems use L as the frequency sampling rate [5], [18].

From the unwrapped phase prospective, the phase spectrum is dependent on the frequency rate, which is often a multiplication of a power of two of the window size $N = 2^p \times L$, where p is a non-negative integer. For conventional sampling at $p = 0$, the unwrapped phase is random to some degree. While for $p \geq 1$, there exists a structure in the unwrapped phases of anechoic and reverberant speech [19], implying a great potential to learn the phase of anechoic speech from that of the reverberant data.

From the above discussions, we can see that the sampling strategy needs to be tailored to reflect the characteristics of traveling echoes in reverberant environments and the properties of the unwrapped phase spectrum. Thus, there is a great need to investigate the effects of sampling in the time and frequency domains on DNN-based dereverberation systems.

IV. Experiments and Result Analysis

The experiments were conducted in a simulated room of dimension 6 by 4 by 3 meters (length by width by height). The positions of the loudspeaker and the microphone were at (2, 3, 1.5) and (4, 1, 2) meters, respectively. Ten RIRs were simulated using an improved image-source (ISM) [20] with reverberation time (RT60) [21] ranging from 0.1 to 1.0 sec, with an increment of 0.1 sec. In order to learn a high-quality DNN model, all 4620 training utterances from the TIMIT set [22] were convolved with the generated RIRs to build a large training set, resulting in about 40 hours of reverberant speech. To test DNN's generalization capability in mismatch conditions, RIRs with RT60 from 0.1 to 1.0 sec, with the increment of 0.05 sec (rather than 0.1 sec) were convolved with 100 randomly selected utterances from the TIMIT test set to construct the test set.

As for acoustic signal processing, all utterances were sampled at a rate of 16 kHz, and the frame length was set to 32 ms (or 512 samples). In addition, PESQ, which has a high correlation with subjective score [11], was used to evaluate the dereverberation results.

Kaldi [23] was used to train DNNs. The DNN configuration was 3 hidden layers, 2048 nodes and 7 frames of input feature expansion for each layer. The number of pre-training epochs for each RBM [24] layer was 1. The learning rate of pre-training was 0.4. As for fine-tuning, the learning rate and the maximum number of epochs were 0.00008 and 30, respectively. The mini-batch size was set to 128. Input and target features of DNN were globally normalized to zero mean and unit variance [18].

A. Time Sampling on Enhanced Speech Quality

To study the frame shift's effects on the DNN-based dereverberation performances, a number of DNNs, whose training and testing utterances were enframed by different frame shift sizes with 512-point DFT computed for each overlapping segment, are presented in Figs. 2, 3 and 4, denoted as “DNN-2ms”, “DNN-4ms”, etc. “Rev” represents unprocessed reverberant speech.

1) Frame Shift = 1/2 Frame Length: We follow the DNN dereverberation results in [5] closely, which utilized a sigmoid output layer and normalized the target features into an unit range of [0, 1], referred as Han’s model. While we proposed to adopt a linear output layer and to globally normalize the target features into zero mean and unit variance. Performance comparisons between Han’s model (“DNN-Han”) and the proposed DNN model (“DNN-proposed”) with conventional time sampling (R = 1/2 L), were given in Fig. 2. When compared with Han’s dereverberation results, our DNN could significantly boost PESQ by 0.31 on the average at all RT60s. Another advantage of the proposed DNN was that it could substantially improve the speech quality in terms of PESQ at all RT60s. However, when compared with reverberant speech at low RT60s below 0.2 sec, the results of Han’s system started to show some considerable PESQ decreases.

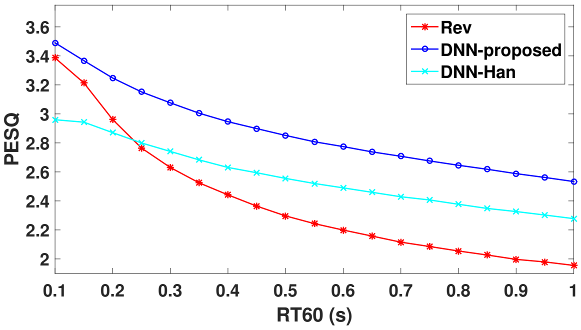

Fig. 2. Average PESQ results on the test set at different RT60s for R = 1/2 L.

2) Frame Shift ≤ 1/2 Frame Length: Fig. 3 presents the improvement of PESQ obtained along all RT60s tested for the conditions of $R \leq$ 1/2 L. “DNN-16ms” (conventional time sampling strategy) was superior to “DNN-2ms” and “DNN-4ms” for RT60 $\geq$ 0.4 sec and 0.6 sec, respectively. In this case, lower frame shifts could not obtain better performances, even at the price of increased computational complexities. To be specific at RT60 = 1 sec, compared with “DNN-16ms”, ’DNN-2ms’ achieved a dramatic PESQ downgrade of 0.19.

3) Frame Shift ≥ 1/2 Frame Length: As shown in Fig. 4, “DNN-24ms” and “DNN-32ms” did not improve the PESQ scores for RT60 $\leq$ 0.2 sec and 0.5 sec, respectively, It was not surprising because dereverberated speech could not be reconstructed exactly at an insufficient time sampling rate that caused aliasing. In addition, “DNN-24ms” improved PESQ scores relative to “DNN-32ms” by 0.38 on the average among all RT60s, illustrating better dereverberation performance could be achieved with less frame shift, if $R >$ 1/2 L.

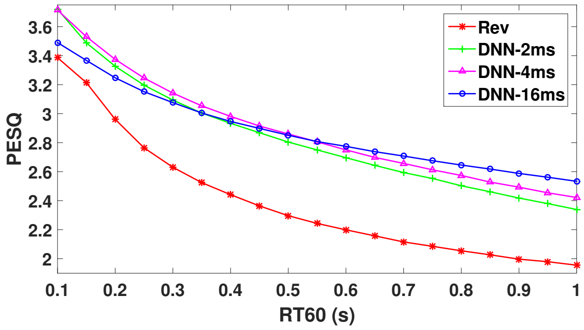

Fig. 3. Average PESQ results on the test set at different RT60s for $R \leq$ 1/2 L.

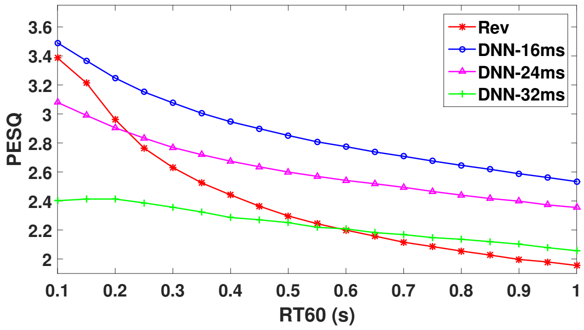

Fig. 4. Average PESQ results on the test set at different RT60s for $R \geq$ 1/2 L.

Similar results were also obtained from frequency-weighted segmental SNR (fwSegSNR) [25] and short-time objective intelligibility (STOI) [26] metrics. With the findings in Figs. 3 and 4, there is a need to adopt an adaptive sampling strategy for DNN-based dereverberation in order to improve the system environmental robustness and performance.

B. Frequency Sampling on Unwrapped Phase

For frequency domain sampling, 512, 1024, 2048, 4096-point DFTs of a short-time audio section were computed, at a fixed frame shift of 16 ms, to present its impact on unwrapped phases of reverberant and anechoic speech.

The unwrapped phase was obtained by the algorithm in [27]. If we assumed that there existed a true unwrapped phase spectrum, each of the above four was an approximation of the true one. Fig. 5 shows that the unwrapped phase was dependent on the DFT size. Specifically, when the DFT size was equal to the frame length (upper left), the unwrapped phases of anechoic and reverberant frame were poor approximations [19]. With more dense frequency sampling rates (upper right, bottom left, and bottom right), the difference between the unwrapped phases of anechoic and reverberant speech behaved similarly, indicating a potential estimation of the true unwrapped phase.

Finally we presented an experiment to highlight the importance of the phase in dereverberation. We combined the

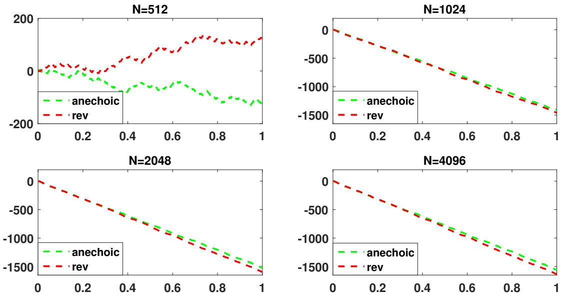

Fig. 5. Unwrapped phase spectra of anechoic and reverberant frames with different DFT lengths (512, 1024, 2048, 4096). The window length is 512 (32 ms at a sampling rate of 16 kHz), and the X-axis denotes the normalized frequency.

TABLE I
AVERAGE PESQ RESULTS ON THE TEST SET AT DIFFERENT RT60S FOR $R = 1/2\ L$

RT60 (s)	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
Rev	3.39	2.96	2.63	2.44	2.30	2.20	2.12	2.05	2.00	1.96
DNN-baseline	3.49	3.25	3.08	2.95	2.85	2.77	2.71	2.65	2.59	2.53
DNN-oracle	3.71	3.65	3.55	3.47	3.38	3.30	3.22	3.15	3.07	3.00

estimated magnitude spectra feature with reverberant (‘DNN-baseline”) and anechoic (“DNN-oracle”) phases to reconstruct the waveforms for $R = 1/2\ L$, although the anechoic speech’s phase could not be obtained in practice. Clearly, in the PESQ comparison in Table I, “DNN-oracle” significantly improved the performance of “DNN-baseline”. Due to the negligible difference between unwrapped phases of anechoic and reverberant speech with a dense sampling in frequency (shown in Fig. 5), there is a great potential for DNN to estimate anechoic speech’s phase from reverberant data, and thus further enhance the DNN-based dereverberation system.

V. CONCLUSION AND FUTURE WORK

In this paper, we investigate the effects of time and frequency sampling rates for STFT and its modifications on the DNN-based dereverberation performances. We show that for time domain sampling, denser rates can not obtain better performances in stronger reverberant environments, even at the price of increased computational complexities. On the other hand for frequency domain sampling, the difference between the unwrapped phases of reverberant and anechoic speech becomes negligible with a dense sampling in frequency. In the future, we will conduct research in developing an adaptive time and frequency sampling strategy to improve the DNN system environmental robustness and performance.

VI. ACKNOWLEDGMENT

The first author was supported by a grant from the China Scholarship Council and this work was partially supported by the National Natural Science Foundation of China (61571344). The authors would like to thank Dr. Yong Xu at University of Science and Technology of China for his valuable suggestions.

REFERENCES

[1] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*. London, UK: Springer, 2010.

[2] M. Wu and D. L. Wang, “A two-stage algorithm for one-microphone reverberant speech enhancement,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 774–784, 2006.

[3] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, “Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 534–545, 2009.

[4] S. Mosayyebpour, H. Sheikhzadeh, T. A. Gulliver, and M. Esmaeili, “Single-microphone LP residual skewness-based inverse filtering of the room impulse response,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1617–1632, 2012.

[5] K. Han, Y. Wang, D. L. Wang *et al.*, “Learning spectral mapping for speech dereverberation and denoising,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 6, pp. 982–992, 2015.

[6] S. T. Neely and J. B. Allen, “Invertibility of a room impulse response,” *J. Acoust. Soc. Amer.*, vol. 66, no. 1, pp. 165–169, 1979.

[7] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “A regression approach to speech enhancement based on deep neural networks,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 7–19, 2015.

[8] P. S. Huang, M. Kim *et al.*, “Deep learning for monaural speech separation,” in *Proc. ICASSP*, 2014, pp. 1562–1566.

[9] K. Li and C.-H. Lee, “A deep neural network approach to speech bandwidth expansion,” in *Proc. ICASSP*, 2015, pp. 4395–4399.

[10] B. Wu, K. Li, M. L. Yang, and C.-H. Lee, “A study on target feature activation and normalization and their impacts on the performance of DNN based speech dereverberation systems,” in *Proc. APSIPA ASC*, 2016.

[11] ITU-T, Rec. P.862, “Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs,” *Int. Telecommun. Union-Telecommun. Stand. Sector*, 2001.

[12] P. Schniter, “Short-time fourier transform,” *Version*, vol. 2, no. 2005, p. 21, 1915.

[13] L. R. Rabiner and R. W. Schafer, *Theory and Application of Digital Signal Processing*. Upper Saddle River, NJ, USA: Prentice Hall, 2010.

[14] J. Benesty, M. M. Sondhi, and Y. Huang, Eds, *Springer Handbook of Speech Processing*. Berlin Heidelberg, Germany: Springer, 2008.

[15] C.-S. Jung, K. J. Han, H. Seo, S. S. Narayanan, and H.-G. Kang, “A variable frame length and rate algorithm based on the spectral kurtosis measure for speaker verification,” in *Proc. Interspeech*, 2010, pp. 2754–2757.

[16] H. Kuttruff, *Room Acoustics*. Oxfordshire, UK: Spon Press, 2009.

[17] R. M. II, *Advanced topics in Shannon sampling and interpolation theory*. Springer Science & Business Media, 2012.

[18] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “An experimental study on speech enhancement based on deep neural networks,” *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, 2014.

[19] A. Rad and T. Virtanen, “Phase spectrum prediction of audio signals,” in *Proc. ISCCSP*, 2012, pp. 1–5.

[20] E. Lehmann and A. Johansson, “Prediction of energy decay in room impulse responses simulated with an image-source model,” *J. Acoust. Soc. Amer.*, vol. 124, no. 1, pp. 269–277, 2008.

[21] W. Sabine, *Collected papers on acoustics*. London, UK: Harvard University Press, 1922.

[22] J. S. Garofolo, “Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database,” NIST, Tech. Rep., 1988.

[23] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, “The Kaldi speech recognition toolkit,” in *Proc. ASRU*, 2011, pp. 1–4.

[24] Y. Bengio, “Learning deep architectures for AI,” *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[25] J. M. Tribolet, P. Noll, B. J. McDermott, and R. E. Crochiere, “A study of complexity and quality of speech waveform coders,” in *Proc. ICASSP*, 1978, pp. 586–590.

[26] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[27] K. Itoh, “Analysis of the phase unwrapping algorithm,” *Applied Optics*, vol. 21, no. 14, p. 2470, 1982.