PERFORMANCE STUDY OF A CONVOLUTIONAL TIME-DOMAIN AUDIO SEPARATION NETWORK FOR REAL-TIME SPEECH DENOISING

Samuel Sonning, Christian Schüldt, Hakan Erdogan, Scott Wisdom

Google Inc.

ABSTRACT

Time-domain audio separation networks based on dilated temporal convolutions have recently been shown to perform very well compared to methods that are based on a time-frequency representation in speech separation tasks, even outperforming an oracle binary time-frequency mask of the speakers. This paper investigates the performance of such a time-domain network (Conv-TasNet) for speech denoising in a real-time setting, comparing various parameter settings. Most importantly, different amounts of lookahead are evaluated and compared to the baseline of a fully causal model. We show that a large part of the increase in performance between a causal and non-causal model is achieved with a lookahead of only 20 milliseconds, demonstrating the usefulness of even small lookaheads for many real-time applications.

***Index Terms*—** Speech enhancement, noise reduction, deep learning, convolutional neural networks, time domain

1. INTRODUCTION

Time-frequency domain representations have been used extensively in traditional noise reduction methods, focusing on estimating and removing the spectral amplitude of noise [1, 2]. Since under reasonable assumptions, the optimal estimate of the clean speech spectral phase is the (noisy) phase of the noisy speech [3], the problem effectively becomes that of estimating the spectral amplitude of the clean speech and using the phase of the noisy speech for reconstruction. A common approach is to use a voice activity detector [4] to control the estimation of the noise spectral amplitude. However, differentiating between highly non-stationary noise and speech is very challenging for traditional methods that rely on long-term statistical estimation of the noise spectrum.

Recent advances in deep learning systems for speech separation have shown to be effective in solving this problem [5, 6, 7, 8, 9]. These methods are typically, as the traditional ones, also based on a time-frequency representation, estimating a multiplicative spectrogram mask to remove noise from the noisy speech mixture. While the time-frequency representation approach is straight-forward and intuitive, it has inherent problems related to phase/magnitude decoupling, and the long time window required to achieve sufficient frequency resolution [10]. To overcome this, Luo and Mesgarani [11] recently proposed a time-domain audio separation network (Conv-TasNet) based on a convolutional encoder followed by temporal dilated convolutional separation layers, an architecture which was shown to significantly outperform previous time-frequency methods as measured by accuracy of separation of speakers in mixed audio – even when compared to an oracle time-frequency ratio mask of the speakers.

Based on Conv-TasNet's success in speech separation, this paper investigates its performance for speech denoising. Specifically, we evaluate a number of nearly causal variants of the network, varying the length of lookahead. Similar studies have been made for time-frequency domain based long short-term memory (LSTM) models, see e.g., [7, 5, 8]. The emphasis of this paper is the performance of the time domain separation network under real-time constraints, including limits to latency, parameter size, I/O bandwidth and computational power. We show that performance scales with length of lookahead, and is especially important for high attenuation in noise-only segments. Also, we observe that using lookaheads of < 10 ms, Conv-TasNet adds audible buzzing artifacts during speech, and frequently leaves sharp artifacts during noise-only segments. Extending the Conv-TasNet lookahead to ≥ 20 ms significantly reduces the artifacts and seems to be a good compromise between low latency and high quality. Moreover, it is concluded that the subjective observations correlate well with objective measures.

2. TIME-DOMAIN AUDIO SEPARATION

In contrast to time-frequency based approaches, Conv-TasNet [10, 11] operates on blocks of time domain samples directly, bypassing the steps of short-time Fourier transform (STFT) and inverse STFT (iSTFT). The model consists of three different steps: encoding, separation, and decoding. The encoder, embodied by a 1-D strided convolution operation followed by a rectified linear unit (ReLU), transforms short overlapping frames with duration of 1.25-5.0 ms of the input signal into a non-negative high-dimensional representation. Separation (in this context denoising, i.e., separating the speech from the noise) is done, similar to the standard STFT/iSTFT time-frequency approach, through the application of a multiplicative mask. In the case of Conv-TasNet, the separation mask is calculated by stacks of dilated 1-D temporal convolutional blocks. Each of these blocks extend the effective time context, backward and forward in time, by a factor equal to the dilation factor of the convolutional layer, which is usually 2. The output of the last block is then passed to a final 1×1-convolutional layer followed by a softmax activation function to estimate mask vectors for the sources (here speech and noise) to be separated. Finally, the decoder reconstructs the waveforms using a 1-D linear deconvolution operation. (The interested reader is referred to [11] for more details).

Other time-domain methods have been recently proposed for audio source separation include U-Net inspired models such as Wave U-Net [12] and SEGAN [13]. In this paper we restrict our attention to Conv-TasNet, since the multiresolution nature of these models make them unsuitable for real-time processing.

Channel-wise layer normalization [10] is used throughout in this paper, with the motivation that this is straight-forward and can be applied in both causal and non-causal cases, as opposed to global layer normalization. The latter can only be used for the non-causal implementation and batch normalization, which imposes a scale dependency between training and testing.

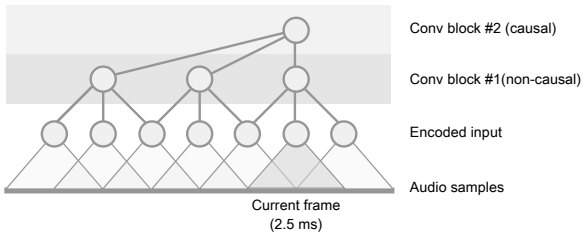

Fig. 1. Illustration of how the non-causal convolutional blocks are used to achieve lookahead in the model.

2.1. Causal models

In telecommunication, low latency is of great importance. ITU-T G.114 [14] suggests an end-to-end one-way ("mouth-to-ear") delay below 150 ms, but also states that highly interactive tasks may be affected by delays below 100 ms. Note that this is the total end-to-end latency, including processing, coding, transmission etc.

Typically, STFT/iSTFT based time-frequency approaches operate on audio frames of a fixed size / time-resolution, where long frames give high frequency resolution yet reduce ability to capture short-time changes. A common choice for frame size is around tens of milliseconds, as that is a time frame where speech can be regarded as fairly stationary [15], and 50-75% overlap.

When it comes to Conv-TasNet, the time context of the dilated convolutional layers extend forward in time as well as backward, implying very high latencies for configurations with many layers. The original Conv-TasNet paper compared such a non-causal configuration to a causal one, where the audio input is time shifted to effectively move the entire time context back in time, allowing causal operation. As having no future information in addition to the current frame has been shown to be detrimental to performance [11], the approach suggested in this paper is to add configurable lookahead by making the lowermost n blocks of the convolutional stack non-causal, while keeping the rest of the stack causal. This is illustrated in Figure 1. The hypothesis is that adding just a few non-causal blocks could provide the network with critical context about the signal to allow better separation, without introducing excessive latency. For convenience we define the lookahead as the future part of the receptive field as seen from the *middle* of the current frame. Note the slight difference from the "signal latency" which is actually larger. In practice, for a window size of W_s ms and a hop size of $W_s/2$ ms, this results in $2^n(W_s/2)$ ms effective lookahead when the kernel size is 3 and dilation factor is 2. For example, using a 2.5 ms window, the lookahead for using n non-causal layers will be 1.25 ms for $n = 0$, 2.5 ms for $n = 1$, 5 ms for $n = 2$, and so on. The latency will then be an additional $W_s/2$ ms, due to the lookahead reference being in the middle of the frame, plus additional sample acquisition/buffering and processing time.

This way of achieving lookahead embedded deep in the network will allow prediction and processing of intricate time based features, and is supported by the experimental results by Luo and Mesgarani [11], which suggests that stacking more convolutional blocks with smaller dilation factors is superior to fewer blocks with larger dilation factors having the same receptive field size.

2.2. Design parameters and constraints

Implementing a network means setting a number of design parameters that affect the network's learning capacity and performance, as well as training time, computational requirements, parameter size and I/O bandwidth. We are interested in exploring performance for real-time applications and for that purpose conduct two sets of experiments. Our first set of experiments focuses on varying lookahead. In our second set of experiments, we try three different ways of further reducing the bandwidth, parameter size and computational requirements. The parameters for the first set of experiments are listed in Table 1, which are values chosen based on the best configuration reported by [11]. For the second set of experiments, we vary the parameters as described in Table 2.

Value	Parameter
2.5	Window length (ms)
256	Number of encoder basis vectors
256	Bottleneck size
512	Number of convolutional channels
3	Kernel size
8	Convolutional blocks per repeat
4	Repeats

Table 1. Parameters used in the first set of experiments.

Window	Blocks / Repeats / Bottleneck	GOps/s	I/O (MB/s)	Params (MB)
2.5 ms	8 / 4 / 256	6.7	26.2	8.39
5.0 ms	8 / 4 / 256	2.9	11.5	7.34
2.5 ms	7 / 4 / 128	2.9	11.5	3.67
2.5 ms	9 / 2 / 128	1.9	7.4	2.36

Table 2. Parameters and requirements for the second set of experiments.

2.3. Model and training

Both the TasNet models and the LSTM model were trained using the same training data, comprising speech samples from LibriTTS [16], mixed with noise samples from Freesound [17], using a normally distributed signal-to-noise ratio (SNR) with mean 5 dB and standard deviation 10 dB, for a total of 134.2 h for the training set, and 6.2 h each for the validation and test sets. The sampling frequency was 16 kHz. For the LSTM model, the STFT of the noisy input signal was calculated, and based on this the compressed magnitude spectrum (exponent 0.3) was passed to the network to infer a mask. The loss function for training was the mean square error between the masked noisy compressed magnitude spectrum and the compressed magnitude spectrum of the clean signal. The Conv-TasNet model is passed noisy time-domain frames, and as in [11], and the training used the negative log SNR loss, using the clean speech as the reference signal.

3. PERFORMANCE EVALUATION

The performance of causal Conv-TasNet with varying lookahead was evaluated for a set of different parameters, and was compared to that of the non-causal Conv-TasNet approach, a causal time-frequency LSTM-based model with 16 ms frame size, and a traditional denoising method here denoted baseline LogMMSE [18]. The LSTM model consists of an input layer of size 129 (i.e., the compressed spectral magnitudes of a single frame), followed by 4 LSTM layers of size 400 each, and two fully-connected layers of size 800 each. The output is a speech mask that is multiplied with the noisy spectrum to achieve a denoised output. For the lookahead parameter of Conv-TasNet, values in a range of 10-80 ms were evaluated, as

Label	Speech energy	Noise energy	SNR
Speech dominant	≥ -60 dBFS	-	≥ 30 dB
Noise dominant	≤ 60 dBFS	≥ -60 dBFS	-
Intermixed	≥ -60 dBFS	-	< 30 dB

Table 3. Labelling of the different evaluation subsets based on speech energy, noise energy and SNR, respectively.

well as a causal variant (1.25 ms lookahead in our terminology, i.e. half the frame) and a non-causal variant (1.28 s lookahead).

3.1. Metrics for evaluation

All evaluated methods (as described in the section above) were evaluated over two different sets of mixed speech/noise audio, using three methods of comparison:

1. Mean SNR improvement (SNRi) over an evaluation subset of the LibriTTS + Freesound dataset (L+F).
2. Scale-invariant signal-to-distortion ratio (SI-SDR) improvement, and scale-invariant signal-to-artifacts ratio (SI-SAR) [19] over a set of files with various types of noise (both ongoing and transient) mixed together with speech.
3. Unofficial human listening tests of the files in item 2 above to subjectively judge quality and determine the characteristic of remaining artifacts.

In the evaluation, we quote SNR improvement (SNRi) over the evaluation set of our training dataset. In addition, we wanted to more thoroughly evaluate the performance over specific SNRs and evaluate the character of artifacts for common types of office noise scenarios. We opted away from evaluating on other common datasets such as CHiME2[20], due to the noise in that dataset being from a noisy living room, including e.g. speech from a television or radio, as well as from people talking in the background, resulting in an undesired ambiguity as to what part of the signal is desired.

We instead decided to record a small test set for a detailed evaluation, including 8 noise tracks of 83 s each, with different categories of office noise, such as keyboard typing, scraping and slamming of coffee cups, paper rustling and fan noise. These tracks were mixed together with clean speech (Harvard Sentences) from four speakers, two male and two female, from ETSI [21]. The set of test files were labelled on a scale of 0.5 s (yielding a total of 1328 labelled segments) with one of three different labels: *speech dominant*, *noise dominant*, and *intermixed sections*. The split between these was done based on the energies and SNR, respectively, as shown in Table 3. The motivation for this is that the employed metrics are sensitive to the input SNR of the segments, with SI-SDR being especially misleading when the speech track is silent or nearly silent, making averaging results across these classes undesired. The resulting mean input SNRs were 10.19 ($\pm$9.30) dB, 35.69 ($\pm$2.86) dB and -36.48 ($\pm$9.93) dB for the mixed, speech dominant and noise dominant subsets, respectively. Furthermore, the relative performance in cases of noise-only, speech-only and mixed speech and noise differ between approaches (as shown in section 4), and has critical perceptual significance, in the sense that e.g. remaining artifacts in noise-only segments will be more clearly heard, without the presence of masking speech.

The manually created test set was evaluated using two objective metrics: SI-SDR improvement (SI-SDRi), and SI-SAR [19] for speech dominant and intermixed categories, and mean power reduction for the noise dominant category. The motivation for using SI-SDRi and SI-SAR is that these measures are both simple and robust.

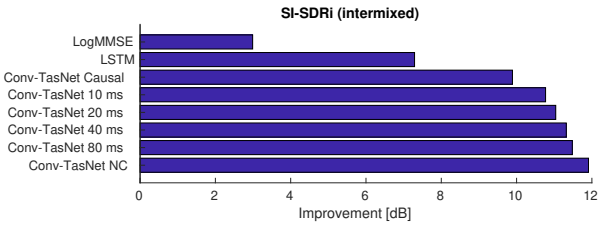

Fig. 2. Comparison of SI-SDRi for the intermixed labels and different lookaheads.

However, when there is no significant speech energy, these measures are less meaningful, hence the use of mean power reduction in these cases. Neither SI-SDRi nor SI-SAR considers scaling mismatch of the reference as an error [19].

4. RESULTS

4.1. Experiment set 1: varying lookahead

Figure 2, Figure 3, and Figure 4 show the performance in terms of SI-SDRi, SI-SAR, and noise reduction, applied to the different signal categories (as described in the previous section) for the different methods over various lookaheads. The performance of the baseline LogMMSE and the time-frequency LSTM are also shown for reference. From the figures, it can be seen that all variants of Conv-TasNet have greatly increased SI-SDRi and noise suppression compared to the LSTM model, which in turn outperforms the classical LogMMSE algorithm. The Conv-TasNet model performs better in these metrics as its lookahead is increased. Interestingly, more than half of the performance difference in SI-SDRi between the causal and non-causal Conv-TasNet variants can be achieved by increasing the lookahead to just 20 ms. Looking at the results for the SI-SAR metric, shown in Figure 3, it can be seen that LogMMSE and the causal Conv-TasNet perform the worst, both achieving lower results than the LSTM model. This is in agreement with our perceptual evaluation, where LogMMSE tends to make speech sound hollow even in relatively high SNR conditions, and causal Conv-TasNet introduces buzzing artifacts. For the LSTM model, speech at times sounds garbled or structurally distorted. Intuitively, this may be an artifact from reusing the noisy phase during signal reconstruction.

When increasing the lookahead, the buzzing artifacts disappear (for lookaheads $>$ 10 ms) and the quality of the denoised speech is discernably better, with the largest jump in quality happening from 1.25 ms (causal) to 10 ms, and from 80 ms to the non-causal variant. These subjective results align with the numbers in Figure 2 and Figure 3, where especially the jump from causal to 10 ms lookahead is notable, both in terms of SI-SAR and SI-SDRi. Perceptually, the 20 ms lookahead model overall sounds better than the LSTM model, which in turn sounds similar to the 10 ms model, but significantly better than the causal variant. This is again in agreement with the objective results; the 20 ms lookahead model clearly outperforms LSTM in terms of SI-SDRi (Figure 2), whereas for SI-SAR the Conv-TasNet versions with lookahead perform much better than Conv-TasNet causal (Figure 3). Interestingly, the SI-SAR results of the causal Conv-TasNet are actually worse than that for the LSTM. It should be noted that SI-SAR was not considered in the original TasNet papers [10, 11], and that the usefulness of separating out SAR from SDR has been questioned [19].

Many speech enhancement researchers rely on full-reference objective metrics like PESQ, and POLQA [22] to evaluate their methods. However, the correlation of these metrics with subjective mean opinion score (MOS) is often not high enough to be reliable for many

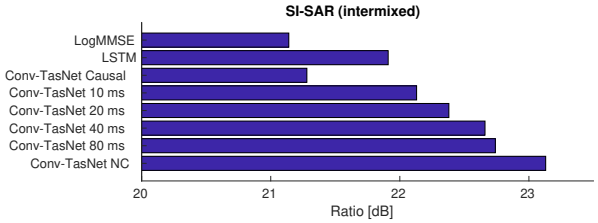

Fig. 3. Comparison of SI-SAR for the intermixed labels and different lookaheads.

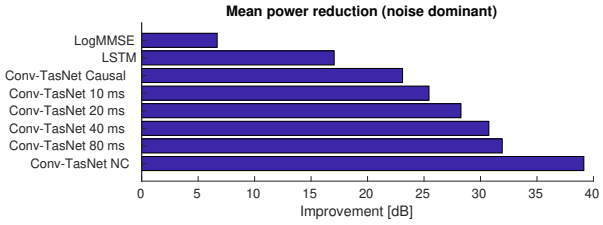

Fig. 4. Noise suppression comparison of segments labelled noise dominant for different lookaheads.

applications [23]. Nevertheless, with this in mind, and taking the estimated scores with a grain of salt, we also compare the denoising methods using ViSQOL [24]. ViSQOL is used for convenience, as the reference MATLAB source code is freely available [25].

The estimated ViSQOL MOS scores agree with the perceived subjective results in that Conv TasNet NC performs the best (3.55) and LogMMSE performs the worst (3.36). The other methods lie in between, in the anticipated order, i.e., longer lookahead giving higher score.

4.2. Experiment set 2: reducing hardware requirements

In addition to the lookahead setting, we explore three other parameter settings, designed to reduce the hardware requirements for real-time execution. The results for these settings are, together with the other results, shown in Table 4. From the resulting SI-SDRi and SNRi, it can be seen, as expected, that reducing the number of convolutional blocks / repeats causes some performance loss. Increasing the window size to 5.0 ms also reduces performance. Both of these results are in line with what is shown in [11], i.e., that with the same network size and separation module configuration, a smaller window size leads to better performance, while with the same window size, fewer blocks or repeats lead to lower performance. In our experiment, these reductions roughly correspond in performance to reducing the lookahead from 20 ms to 10 ms, with matching perceptual evaluation.

Window	Lookahead	Blocks & repeats	SI-SDRi (intermixed)	SNRi (L+F)
2.5 ms	1.25 ms	8/4	9.89 dB	13.16 dB
2.5 ms	10 ms	8/4	10.77 dB	13.49 dB
2.5 ms	20 ms	8/4	11.04 dB	13.83 dB
2.5 ms	40 ms	8/4	11.32 dB	14.06 dB
2.5 ms	80 ms	8/4	11.48 dB	14.32 dB
2.5 ms	1280 ms	8/4	11.91 dB	15.14 dB
2.5 ms	20 ms	7/4	10.71 dB	13.48 dB
2.5 ms	20 ms	9/2	10.58 dB	13.32 dB
5.0 ms	20 ms	8/4	10.38 dB	13.54 dB

Table 4. Parameters used in the second set of experiments, together with the resulting SI-SDRi and SNRi. All other parameters were as described in Table 1.

5. DISCUSSION

Although not directly comparable (due to difference in evaluation data set and model parameters), our results for SI-SDRi in the mixed case are similar in magnitude to those reported in [11] for various configurations of the causal model with 2.5 ms window size. We, however, see a smaller difference between the non-causal and causal versions than reported there: around 2 dB according to both our measures, compared to 3.5 dB in [11]. It should be noted, however, that the perceptual difference of those 2 dB is very large; in general, we have found that in evaluation of denoised speech, even very small SI-SDR increases, on the scale of less than 0.5 dB, can lead to clear perceptual improvements, mainly due to the reduction of remaining audible artifacts.

From our results we can clearly see the importance of using future audio context for the denoising task. We can also observe that although it may be helpful to provide context of up to 80 ms or more into the future, by that point we have already included a majority of the important information for discriminating between speech and noise. These findings are not surprising considering that speech is highly auto-correlated over short time-frames, and future information should thus be useful for increasing confidence that a certain part of a signal is speech. Our results align with [7], but contrast with those in [8], both exploring lookahead for time-frequency LSTM, where no significant difference was found between the denoising performance of a bi-directional recurrent network, a fully causal unidirectional variant, and a unidirectional variant with 0-200 ms lookahead. (It should though be noted that [7], [8] use different datasets, which might, at least partially, explain the difference in results.)

The most significant difference between different lookaheads in the Conv-TasNet model appears to be suppression of noise in noise-dominant parts of the signal (Figure 4). The suppression performance on these segments is remarkably low for the models with short lookahead, given that it should be straightforward to simply null the signal if there are no indications of speech. One explanation for this could be the training dataset and loss function used; the dataset contains few segments with only noise, compared to the number of segments dominated by speech, and the log SNR loss function isn't as meaningful when there is no or little speech energy in the target signal. Based on this explanation, an alternative way of boosting performance during noise-only segments may be to alter the loss function to better account for suppression of noise when there is no speech present.

6. CONCLUSION

The performance of time-domain audio separation networks based on dilated temporal convolutions has been studied for different lookaheads. It has been shown that the performance of the considered causal versions with lookahead ≥ 10 ms surpass the benchmark time-frequency based LSTM approach in SI-SDRi, SI-SAR, SNRi as well as subjective audio quality. As expected, the performance increases with increased lookahead, with the largest jumps in quality happening between 1.25 ms and 10 ms, and between 80 ms and the non-causal variant. As such, we believe that lookaheads between 10 ms and 40 ms are especially useful for many real-time denoising applications, as they significantly increase quality over the causal case, while keeping latency within an acceptable span.

7. REFERENCES

[1] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen, *Noise reduction in speech processing*, vol. 2, Springer Science & Business Media, 2009.

[2] Yi Hu and Philipos C Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech communication*, vol. 49, no. 7-8, pp. 588–601, 2007.

[3] Yariv Ephraim and David Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[4] Javier Ramirez, Juan Manuel Górriz, and José Carlos Segura, "Voice activity detection. fundamentals and speech recognition system robustness," in *Robust speech recognition and understanding*. InTech, 2007.

[5] Hakan Erdogan, John R Hershey, Shinji Watanabe, and Jonathan Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 708–712.

[6] Donald S Williamson, Yuxuan Wang, and DeLiang Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 3, pp. 483–492, 2016.

[7] Gordon Wichern and Alexey Lukin, "Low-latency approximation of bidirectional recurrent networks for speech denoising.," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 66–70.

[8] Kevin Wilson, Michael Chinen, Jeremy Thorpe, Brian Patton, John Hershey, Rif A Saurous, Jan Skoglund, and Richard F Lyon, "Exploring tradeoffs in models for low-latency speech enhancement," in *Proceedings of the 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2018, pp. 366–370.

[9] Scott Wisdom, John R. Hershey, Kevin Wilson, Jeremy Thorpe, Michael Chinen, Brian Patton, and Rif A. Saurous, "Differentiable consistency constraints for improved deep speech enhancement," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019.

[10] Yi Luo and Nima Mesgarani, "TasNet: time-domain audio separation network for real-time, single-channel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 696–700.

[11] Yi Luo and Nima Mesgarani, "TasNet: surpassing ideal time-frequency masking for speech separation," *arXiv preprint arXiv:1809.07454*, 2018.

[12] Daniel Stoller, Sebastian Ewert, and Simon Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[13] Santiago Pascual, Antonio Bonafonte, and Joan Serra, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[14] ITU-T, "G.114 One-way transmission time," *G-Series recommendations*, 2003.

[15] Boaz Porat, *Digital processing of random signals: theory and methods*, Courier Dover Publications, 2008.

[16] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu, "LibriTTS: a corpus derived from LibriSpeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[17] Freesound, `https://freesound.org`.

[18] Y Ephraim and David Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, pp. 443 – 445, 05 1985.

[19] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey, "SDR - half-baked or well done?," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2019.

[20] Emmanuel Vincent, Jon Barker, Shinji Watanabe, Jonathan Le Roux, Francesco Nesta, and Marco Matassoni, "The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 126–130.

[21] ETSI, "Speech and multimedia Transmission Quality (STQ); Speech quality in the presence of background noise: Objective test methods for super-wideband and fullband terminals," *ETSI TS 103 281*, 2017.

[22] John G Beerends, Christian Schmidmer, Jens Berger, Matthias Obermann, Raphael Ullmann, Joachim Pomy, and Michael Keyhl, "Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement part I - Temporal alignment," *Journal of the Audio Engineering Society*, vol. 61, no. 6, pp. 366–384, 2013.

[23] Chandan KA Reddy, Ebrahim Beyrami, Jamie Pool, Ross Cutler, Sriram Srinivasan, and Johannes Gehrke, "A scalable noisy speech dataset and online subjective test framework," *arXiv preprint arXiv:1909.08050*, 2019.

[24] Andrew Hines, Jan Skoglund, Anil C Kokaram, and Naomi Harte, "ViSQOL: an objective speech quality model," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 13, 2015.

[25] ViSQOL Software, `http://sigmedia.tv/tools`.