Hands-free Speech Communication Using Deep Neural Networks

Amir Ivry

Andrew and Erna Viterbi Faculty of Electrical and Computer Engineering
Technion – Israel Institute of Technology, Technion City, Haifa 3200003, Israel
sivry@campus.technion.ac.il

Abstract

Speech processing in challenging acoustic conditions has been an active area of research for decades. It has been shown that acoustic environments of high levels of noise and echo, transient interference, reverberations, and degraded audio measurements, which are practically unavoidable, lead to deteriorated performance in most speech-based systems. This research examines various speech signal representations in both time and time-frequency domains and introduces novel deep learning architectures. In particular, we aim to develop real-time and low-resources implementations that can be embedded into speech-based hands-free communication platforms, and to apply them into integrated frameworks in these acoustic environments, e.g., for voice activity detection, residual echo suppression, nonlinear acoustic echo cancellation, and speech enhancement.

Index Terms: Hands-free speech communication, real-time and low-resources implementations, deep learning

1. Introduction

The problem of speech processing in challenging acoustic conditions has been an active area of research for decades. It has been shown that low signal-to-noise-ratio (SNR) and signal-to-echo-ratio (SER) levels, transient interference, reverberations, and degraded and distorted measurements, which are unavoidable in real-life scenarios, lead to deteriorated performance in most speech-based systems. Commonly, speech is represented in the time-frequency domain using variations of the short-time Fourier transform (STFT) [1]. In most deep learning approaches that attracted recent research efforts, the STFT is fed as a sequential time frame representation into artificial neural networks or into feedback-based recurrent neural networks, or alternatively as a sequence of images into convolutional-oriented neural networks. These sub-optimal solutions frequently pose a system latency that exceeds mobile communication standards, and demand high computational resources that are often impractical for embedding on common processors for mobile communication platforms. These considerations may render existing deep learning-based speech processing systems inadequate for real-time hands-free communication.

In our preliminary work, we focused on three speech-based applications: voice activity detection (VAD), residual echo suppression (RES), and nonlinear acoustic echo cancellation (NLAEC). In each of these studies, we aimed to achieve leading results in real-life acoustic setups while constructing low-latency and low-resources neural network-based systems that are adequate for real-time on-device communication platforms. We first introduced a VAD system [2, 3] that utilized the compact mel scale time-frequency representation [4], which is a compression of the STFT representation. The spectral features were fed into a small-scale fully-connected encoder-decoder-based neural network. To address RES, we integrated the efficient depth-wise-separable convolution technique on the UNet convolutional deep learning model [5] for reconstruction of speech from highly degraded measurements [6]. NLAEC was also addressed by exploiting low-scale feedback-based neural networks that were directly fed with the waveform representation of the speech signals [7]. Recently, we introduced two objective metrics we developed to separately quantify the desired-speech maintained level and residual-echo suppression level during double-talk periods [8].

In this research, we aim to examine real-valued and waveform-based speech signal representations, to develop a dedicated speech-based framework for these representations, and to extend these frameworks to additional speech processing applications for hands-free communication. Specifically, we target speech enhancement in transient noisy environment, speech intelligibility enhancement for eavesdropping in unknown acoustic environments, and acoustic fencing appliances.

2. Preliminary Results

We addressed VAD in reverberant acoustic environments of transients and stationary noises [2], [3], which often occur in real-life scenarios. We exploited unique spatial patterns of speech and non-speech audio frames by independently learning their underlying geometric structure. This process was done through a deep encoder–decoder-based neural network architecture that involves an encoder that maps spectral features with temporal information to their low-dimensional representations, which are generated by applying the diffusion maps method [9]. The encoder feeds a decoder that maps the embedded data back into the high dimensional space. A deep neural network, which is trained to separate speech from non-speech frames, was obtained by concatenating the decoder to the encoder, resembling the known diffusion nets architecture [10]. Experimental results showed enhanced performance compared to competing methods in both accuracy, robustness, and generalization ability.

Next, we proposed an RES method based on a UNet neural network that directly maps the outputs of a linear AEC system to the desired signal in the spectral domain [6]. This system embeds a novel design parameter we developed that allows a tunable tradeoff between the desired-signal distortion and residual echo suppression in double-talk scenarios. Experiments are conducted with 161 h of data from the Microsoft AEC-challenge database [11] and from real independent recordings. We demonstrated the superiority of the proposed system in real-life conditions in terms of regarding echo suppression and desired-signal distortion, generalization to various environments, and robustness to high echo levels.

We also developed an NLAEC system [7], which aims to model the echo path from the far-end signal to the near-end microphone in two parts. Inspired by the physical behavior of modern hands-free devices [12], [13], [14], [15], we first introduced a novel neural network architecture that is specifically designed to model the nonlinear distortions these devices induce between receiving and playing the far-end signal. To account for variations between devices, we constructed this network with trainable memory length and nonlinear activation functions that are not parameterized in advance, but are rather optimized during the training stage using the training data. Second, the network was succeeded by a standard adaptive linear filter that constantly tracks the echo path between the loudspeaker output and the microphone. During training, the network and filter are jointly optimized to learn the network parameters. Using 280 h of real and synthetic data [16], experiments show advantageous performance compared to competition in real-life setups.

The recent deep noise suppression mean opinion score (DNSMOS) metric was shown to estimate human ratings with great accuracy [17]. The signal-to-distortion ratio (SDR) metric is widely used to evaluate RES systems by estimating speech quality during double-talk [18]. However, since the SDR is affected by both speech distortion and residual-echo presence, it does not correlate well with human ratings according to the DNSMOS. To address that, we introduced two objective metrics to separately quantify the desired-speech maintained level (DSML) and residual-echo suppression level (RESL) during double-talk [8]. These metrics are evaluated using our deep learning-based RES-system with a tunable design parameter [6]. Using 280 h of real and simulated data [16], we showed that the DSML and RESL correlate well with the DNSMOS with high generalization to various setups. Also, we empirically investigated the relation between tuning the RES-system design parameter and the DSML-RESL tradeoff it creates and offered a practical design scheme for dynamic system requirements.

These systems meet the standard timing requirements of hands-free communication [19] with maximal system latency of 40 ms on a standard neural processor, e.g., the NDP120 by Syntiant$^{\text{TM}}$ [20]. Also, their computational requirements reach maximal amount of 10 Mega-bytes of overall required memory and 1.6 Giga floating-point operations per second (Gflops), which is adequate for integration on common mobile devices.

3. Future Research Objectives and Expected Significance

We aim to decompose the speech waveform signal into its frequency sub-bands using a real-valued transform, in contrast to the common complex-valued representation applied by the STFT and its modifications. This transform can enable a utilization of waveform-based deep learning models, and in certain cases lead to improved performance compared to their STFT-based counterparts. For instance, feedback-based neural networks that are specifically built for time sequence analysis and were shown successful for speech analysis can be applied efficiently using this representation, e.g., [21], [22], [23], [24]. Also, preservation of phase information is achieved, in contrast to STFT-based methods that usually introduce mismatch between the reconstructed amplitude and original phase information. In addition, every sub-band is associated with a lower sample frequency than the original signal, which may reduce the computational complexity and lower the inference time of the system. We already established an API to convert speech waveform into its sub-bands using real-valued signal representation, and future research will involve applying this representation into a waveform-based speech processing framework.

Equipped with a sub-band decomposition of the speech signal, in the next research stage we plan to respectively decompose existing speech-based systems into smaller and more efficient sub-systems. Nowadays, waveform architectures are fed with the complete spectrum of speech signals that often demands high-resources consumption for high-quality modeling, which is not optimal for real-time usage. We aim to process each sub-band representation of the speech signal separately and independently by a smaller waveform-based architecture, and merge their outcomes. We hope that each sub-system will require a small computational load that is reasonable for embedding on real-time mobile communication platforms.

Today, speech enhancement systems are highly desirable for various low-power hands-free communications platforms, such as smartphones, smart speakers, wearable devices, smart homes, IoT endpoints, and more. Speech enhancement resembles our previous studies of residual echo suppression and nonlinear acoustic echo cancellation, since in both cases speech should be recovered from degraded measurements. Speech enhancement also draws similarity to our voice activity detection study that detected speech in transient noisy and reverberant environments. Thus, we aim to project the concepts we already successfully applied in previous systems to speech enhancement. To comprehend various real-life scenarios, we plan to employ a new open source database that contains hundreds of hours of recordings in real-life acoustic conditions [25].

We also aim to address speech intelligibility enhancement for eavesdropping using hidden microphone recordings in unknown acoustic environments. These may include ones with strong reverberations, echoes, and interference, and speech not directed at the microphone reception area. Achieving success in the proposed research is valuable in several aspects. First, creating a low-power speech intelligibility enhancement system and embedding it into stand-alone eavesdropping devices. Second, exploiting narrow-band transmission that is cheap, long-range, and low-power consuming, and extending the longevity of the battery-supplied device. Third, allowing more rapid and improved data inference by the end-user, i.e., the listener. Forth, reducing cost spent on human trainings that include big data collection. And fifth, achieving enhanced performance of following speech recognition algorithms. Similarly to the planned speech enhancement research, we will examine our existing deep learning systems and their performance on this task.

Acoustic fencing aims at separating speakers by their physical locations in a room using a microphone array [26]. Achieving success in this research can benefit many speech-based applications. For instance, it may improve speech enhancement of a speaker located in a certain region by attenuating speech sources that are located in other regions in the room. That and more, it may enhance succeeding speech-based systems, e.g., for direction estimation, speaker recognition, and speech recognition. Another on-demand application nowadays is automatic transcription of conference meetings. By setting acoustic fences that isolate speakers located in different regions in the room, more accurate transcription results can potentially be obtained compared with existing methods.

4. Acknowledgments

The author would like to thank Prof. Israel Cohen and Dr. Baruch Berdugo, his academic supervisors for PhD degree since 2018, for their valuable advisement. This research was supported by the Pazy Research Foundation and the ISF-NSFC joint research program (grant No. 2514/17). The author thanks Stem Audio for providing equipment and technical guidance.

5. References

[1] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transacitons on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[2] A. Ivry, B. Berdugo, and I. Cohen, "Voice activity detection for transient noisy environment based on diffusion nets," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 254–264, 2019.

[3] A. Ivry, I. Cohen, and B. Berdugo, "Evaluation of deep-learning-based voice activity detectors and room impulse response models in reverberant environments," in *Proc. ICASSP*. IEEE, May 2020, pp. 406–410.

[4] S. Umesh, L. Cohen, and D. Nelson, "Fitting the mel scale," in *Proc. ICASSP*, vol. 1. IEEE, 1999, pp. 217–220.

[5] P. Gadosey, Y. Li, E. A. Agyekum, T. Zhang, Z. Liu *et al.*, "SD-UNet: Stripping down U-Net for Segmentation of Biomedical Images on Platforms with Low Computational Budgets," *Diagnostics*, vol. 10, no. 2, p. 110, 2020.

[6] A. Ivry, I. Cohen, and B. Berdugo, "Deep residual echo suppression with a tunable tradeoff between signal distortion and echo suppression," in *Proc. ICASSP*. IEEE, Jun. 2021.

[7] ——, "Nonlinear acoustic echo cancellation with deep learning," in *Proc. Interspeech*. IEEE, Sep. 2021.

[8] ——, "Objective metrics to evaluate residual-echo suppression during double-talk," in *Proc. WASPAA*. IEEE, Oct. 2021.

[9] R. Coifman and S. Lafon, "Diffusion maps," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 5–30, 2006.

[10] G. Mishne, U. Shaham, A. Cloninger, and I. Cohen, "Diffusion nets," *Applied and Computational Harmonic Analysis*, 2017.

[11] K. Sridhar, R. Cutler, A. Saabas, T. Parnamaa, M. Loide, H. Gamper *et al.*, "ICASSP 2021 acoustic echo cancellation challenge: Datasets, testing framework, and results," in *Proc. ICASSP*. IEEE, 2021, pp. 151–155.

[12] A. Dobrucki, "Nonlinear distortions in electroacoustic devices," *Archives of Acoustics*, vol. 36, no. 2, pp. 437–460, 2011.

[13] W. Klippel, "Loudspeaker nonlinearities–causes, parameters, symptoms," in *Audio Engineering Society Convention 119*. Audio Engineering Society, 2005.

[14] R. Ravaud, G. Lemarquand, T. Roussel, and V. Lemarquand, "Ranking of the nonlinearities of electrodynamic loudspeakers," *Archives of Acoustics*, vol. 35, no. 1, pp. 49–66, 2010.

[15] M. Soria-Rodríguez, M. Gabbouj, N. Zacharov, M. Hamalainen, and K. Koivuniemi, "Modeling and real-time auralization of electrodynamic loudspeaker non-linearities," in *Proc. ICASSP*, vol. 4. IEEE, 2004, pp. 81–84.

[16] R. Cutler, A. Saabas, T. Parnamaa, M. Loide, S. Sootla, M. Purin *et al.*, "Interspeech 2021 acoustic echo cancellation challenge," in *Proc. Interspeech*, 2021.

[17] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," *arXiv:2010.15258*, 2020.

[18] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[19] *ETSI ES 202 740: Speech and multimedia Transmission Quality (STQ); Transmission requirements for wideband VoIP loudspeaking and handsfree terminals from a QoS perspective as perceived by the user*, ETSI Std., 2016.

[20] "NDP120 Syntiant™ Neural Processor," https://www.syntiant.com/ndp120, 2021.

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

[22] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *Proc. ICASSP*. IEEE, 2018, pp. 696–700.

[23] ——, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[24] E. Grais, D. Ward, and M. Plumbley, "Raw multi-channel audio source separation using multi-resolution convolutional auto-encoders," in *Proc. EUSIPCO*. IEEE, 2018, pp. 1577–1581.

[25] K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey *et al.*, "The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," *preprint arXiv:2005.13981*, 2020.

[26] S. E. Chazan, H. Hammer, G. Hazan, J. Goldberger, and S. Gannot, "Multi-microphone speaker separation based on deep DOA estimation," in *Proc. EUSIPCO*. IEEE, 2019, pp. 1–5.