AI-based Speech Enhancement for Hearing Aids : A State of the art and study of the perspectives

Jean-Marie Lemercier

***Abstract*—Hearing-Impaired patients are getting more numerous (5 % of the world population today, and that proportion is expected to double by 2050 [1]). This can be accounted for by demographic reasons - an ageing population with natural degeneration of sensory cells - or external reasons – young population being more exposed to noise in recreational and work environments, or infectious diseases affecting foetuses. However, there is also hope for better prevention and treatments as the quality of diagnosis and Hearing Aids solutions is increasing. Particularly in the last decade, breakthrough techniques involving Artificial Intelligence and more specifically Deep Neural Networks (DNN) have provided Speech processing fields with powerful tools that can be used jointly with non-learning methods like Beamforming, Binaural processing or Auditory Scene Analysis to name a few.**

I. Introduction

This study presents an overview of the state-of-the-art methods for AI-based speech processing, implemented for some in hearing aids. Personal views will also be expressed in the second section on techniques which could be considered promising for this field.

II. State-of-the-Art DNN-based speech enhancement algorithms

Speech processing algorithms can be separated into three classes corresponding to specific objectives: *(i)* denoising/speech enhancement, *(ii)* dereverberation and *(iii)* speaker separation. All three will be investigated as we can see some of these targets can jointly be optimized.

A. Time-Frequency domain methods

The most researched and implemented algorithms act on the spectral magnitude in the Time-Frequency (**T-F**) domain, meaning a Short-Term Fourier Transform (**STFT**)is employed to perform a time-frequency analysis of the input samples before doing any further processing or training. The waveform is then reconstructed at the output of the algorithm by using inverse STFT and the noisy input phase. Most *Speech Enhancement* AI-based algorithms train Deep Neural Networks to generate **T-F masks**, that is, to learn from both noisy and clean input samples how to match target time-frequency maps of relative power between noisy and clean signal (examples: Ideal Binary Mask (**IBM**), Ideal Ratio Mask (**IRM**), Spectral Magnitude Mask). Another common target strategy is **spectral mapping**, that is, directly estimating the clean speech spectral power as a time-frequency map (examples: Target Magnitude Spectrum, Gammatone Frequency Magnitude Spectrum).

One of the first DNN-based algorithms is described in [2]: Input samples are passed through a Gammatone Filter bank (which reproduces the log-spaced-frequency-centred filter structure of the cochlea) to produce 64 subband-spectrograms. Separated subband-DNNs train on these subband-spectrograms to extract features and learn to map them on subband-IBMs. The IBMs are then concatenated to fit to the training target, being a fullband IBM. This subband approach was proved by [4] to perform better than simply using a fullband DNN.

Since then, many various studies have brought improvements to this monaural T-F masking approach (proved by [5] to perform better than spectral mapping for denoising purposes). [6] explores the interest of using several training targets by combining T-F masks, [7] has designed a new monaural feature (the most classical ones being Mel Frequency Cepstral Coefficients (**MFCC**) and Gammatone Frequency Cepstral Coefficients (**GFCC**) [3]) called Multi-Resolution Cochleagram (**MRCG**), to gain several levels of context information.

In terms of *Speech Dereverberation*, [10] have presented a DNN approach on the problem by training a network to map reverberant cochleagrams (Gammatone Frequency features) to clean (anechoic) cochleagrams. (Similarly, [5] proved that spectral mapping was more indicated than T-F masking for dereverberating purposes). In 2017, [11] showed that the reverberation time T_{60} had an influence on the optimal frame size and shift for the STFT analysis: in that view, the authors proposed a T_{60}-aware dereverberation DNN algorithm. The weights learnt during training are used during dereverberation stage on features, whose extraction is parametered by an estimation of the T_{60} (tracked on input reverberant speech).

In separating target speaker from interferences, i.e. *Speech separation*, DNN-based algorithms have imposed as standards, showing good results compared to deterministic approaches like Computational Auditory Scene Analysis (CASA), although they use some of the same ideas like clustering T-F bins around source-specific centroïds and could therefore be combined to form an interesting solution. One has to discriminate those algorithms between speaker-dependent and speaker-independent models, the first class being models that must use the same interference speakers for training and testing. A good early example of those models is proposed in [18], where specific layers are used for estimating speakers in parallel with discriminative layers, resulting in a loss function composed of a weighted sum of estimation loss (aiming at producing a good estimation of each speaker) and discrimination loss (aiming at differentiating each speaker estimation from the

other speaker's reference).

However, a realistic Hearing Aid implementation of speech separation algorithms would demand speaker (and target) model independence. To achieve that, two major algorithms have been proposed: Deep Clustering ([14]) (and its variant Deep Attractor Network (**DANet**)) and Permutation-Invariant Transforms ([15]). In [14], Deep Clustering is described as a speaker-independent model, first training a DNN to map T-F units to embedding vectors and then using an unsupervised clustering method (K-Means) to separate the T-F units according to the speakers. DANet takes a step further by bringing the clustering method into the training stage, forming and updating "attractor speakers" with a speaker assignment stage ([16]). Permutation-Invariant Transforms on the other hand, combine outputs of DNNs with different loss functions (related to the different target speakers) and optimize the global sum of permutations of these losses in order to gain speaker-independence. Some methods, like the **ODANet** [17] combine both approaches to achieve better results.

B. Multi-channel methods

Interesting approaches of spatially aware Neural Networks have been proposed, using binaural and array processing methods to add spatial features into the T-F masking training process.

One of the first papers mentioning such technique for *Speaker Separation* would be [8], in which the authors use the left and right channels to produce two GFCC representations, and compute the Interaural Intensity Difference (**IID**) and Interaural Time Difference (**ITD**) for each T-F unit. They use these features in addition with the GFCC map from left-channel clean input to train subband-DNNs and produce a concatenated, fullband IBM mask. This method is shown to generalize well for different spatial and reverberant configurations.

Array microphones can also be used in conjonction with DNN by learning the parameters for a beamformer performing *Speech Enhancement*. In [9], a Minimum Variance Distortionless Response (**MVDR**) beamformer is designed by learning the Noise and Signal Covariance matrixes through a fusion of different IBMs. Each microphone generates a IBM through a bidirectional LSTM network based on the recorded data, and the different masks are combined into one IBM by a median operation [3].

C. Time-domain methods

The issues with Time-Frequency methods are that only spectral magnitude is affected by those algorithms, and the phase is reconstructed in the output by taking the noisy input phase. This phase-mismatch can create important artifacts degrading the speech quality and intelligibility. In addition to that, the use of STFT and iSTFT as pre-processing and post-processing parts of an algorithm necessarily introduces additional aliasing. Furthermore, time window length has to be quite important to achieve good resolution, which makes another meaningful argument in favour of time-domain algorithms because the frame length is supposed to be inferior to 20ms for speech enhancement typically.

If originally DNNs were ill-designed to enhance speech at the waveform level, the emergence of LSTM Recurrent Neural Networks, and more recently Fully Convolutional Neural Networks (or **TCN**, Temporal Convolutional Networks) helped designing time-domain efficient algorithms. The first class helps keep residues of information in the network to gain a "historic" dimension in the training. [19] on the other hand motivates the use of TCN by claiming Convolutional Layers act as feature extractors or filters, which was already known to be effective in the time-domain, but they also observed that removing the Fully-Connected layers of the conventional CNN architecture helped decorrelating the mapping of low and high frequencies of the waveform [3].

Finally, a new state of the art has emerged thanks to temporal methods: **TasNet** [20] is a *Speech Separation* algorithm using a LSTM separator module for producing temporal masks, encased in a encoder-decoder framework. A recent variant of it is **Conv-TasNet** [21] which combines the previous approaches by replacing the LSTM module by a TCN network for temporal separation.

Simultaneously, a *Speech Enhancement* time-based method has been proposed using Relativistic Generative Adversarial Networks (**RGANs**) in [22]. If GANs were not particularly popular in Speech Enhancement applications so far [3], it was mainly because they are very hard to train efficiently. In [22], the use of a relative loss function for comparing noisy and clean speech, as well as the introduction of gradient penalty, help stabilize the training and achieve interesting performances.

III. Recently developed axes of interest for future AI-related research

Several techniques have been introduced in the forehand section, and a few promising axes can be highlighted:

Combination of spatial estimation methods and DNNs could become very interesting for Hearing Aids implementation, since the binaural approach has proved essential and even external microphones solutions have been investigated [25] [26] to overcome the issue of limited number of microphones in Hearing Aids.

Good feature design could be instrumental in capturing the underlying structure of the auditory periphery system into the DNNs technology: one could try to associate automatic feature extraction with CNNs, and bio-inspired ones like GFCC, MRCG [7]. Within the same paradigm, making progress can be achieved by merging domains: GANs are the combined product of Game Theory and Neural Networks, and ResNets (Networks with skip connections) were issued from a deep observation of pyramidal cortical cells in conjonction with Network design. Finally, making a point of finding the best probabilistic models and distributions for the chosen subject can make a true difference: in [13], a semi-supervised Variational Auto-Encoder (**VAE**) is used for a speech model

in conjonction with an unsupervised noise model based on α-stable distributions. The latter are heavy-tailed distributions alleged to be more tailored to fit audio signals than the usual Gaussian assumption [3]. In definitive, I believe a true preliminary study of the precise subject of speech enhancement is needed to come up with the best features, metrics and assumptions, thus adding real research value, rather than just making use of regular models that have been used thousands of times.

Another pragmatically important issue is **reducing computational complexity** of DNN-based algorithms, as we want our methods to be implementable in real-time Hearing Aids solutions. That issue has been broached in [17] where an online solution is proposed for speaker-independent speech separation (ODANet). One of the possible manoeuver to achieve real-time implementation for a DNN-based algorithm would be to focus on pre-training.

Time-domain advantages have been mentioned before, and end-to-end speech processing should be one of the prime points of interest if the industry wishes to propose true audio quality. In that view, recent powerful tools like GANs, LSTMs or TCNs could be interesting to investigate, along with temporal feature design.

Finally, **unifying models** to treat objectives jointly is to me the key point in having true informed methods for Speech Processing that capture the global complexity of the speech situation. For instance, a recent study [23] tackles Dereverberation and Speech Separation with the same algorithm by proposing a unified formalism taking into account the Convolutive structure of reverberant speech in the framework of Multichannel VAE. Even in non-learning methods, this unification paradigm is investigated: [27] proposes to link an auditory attention estimation with EEG signals to update weights in an Adaptive Binaural Beamformer.

IV. Conclusion

Deep Neural Networks offer true potential for implementation of quality-Speech Processing algorithms in Hearing Aids. This technology's ability to perform automatic feature extraction as well as generalize excellent performances to different spatial configurations, noise-types or speakers is unique. It therefore provides hope for proper analysis and treatment of real-life situations for the Hearing Impaired, if properly conceived, developed and combined with existing technologies.

References

[1] World Health Organization
2020 WHO fact sheet on Deafness and Hearing Loss
https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

[2] Y. Wang and D.L. Wang
Towards scaling up classification-based speech separation
IEEE Trans. Audio Speech Lang. Proc., vol. 21, pp. 1381-1390, 2013.

[3] DeLiang Wang, Fellow, IEEE, and Jitong Chen
Supervised Speech Separation Based on Deep Learning: An Overview
Ohio State University, Columbus, 2018.

[4] J. Du and Y. Xu
Hierarchical deep neural network for multivariate regresss
Pattern Recognition, vol. 63, pp. 149-157, 2017.

[5] Y. Zhao, Z.-Q. Wang, and D.L. Wang
A two-stage algorithm for noisy and reverberant speech enhancement
in Proceedings of ICASSP, pp. 5580-5584, 2017.

[6] X. Zhang, H. Zhang, S. Nie, G. Gao, and W. Liu
A pairwise algorithm using the deep stacking network for speech separation and pitch estimation
IEEE/ACM Trans. Audio Speech Lang. Proc., vol. 24, pp. 1066-1078, 2016.

[7] J. Chen, Y. Wang, and D.L. Wang
A feature study for classification-based speech separation at low signal-to-noise ratios
IEEE/ACM Trans. Audio Speech Lang. Proc., vol. 22, pp. 1993-2002, 2014.

[8] J. Heymann, L. Drude, and R. Haeb-Umbach
Neural network based spectral mask estimation for acoustic beamforming
in Proceedings of ICASSP, pp. 196-200, 2016.

[9] H. Erdogan, J.R. Hershey, S. Watanabe, M. Mandel, and J.L. Roux
Improved MVDR beamforming using single-channel mask prediction networks in Proceedings of Interspeech, pp. 1981-1985, 2016.

[10] K. Han, Y. Wang, and D.L. Wang
Learning spectral mapping for speech dereverebaration
in Proceedings of ICASSP, pp. 4661-4665, 2014.

[11] B. Wu, K. Li, M. Yang, and C.-H. Lee
A reverberation-time-aware approach to speech dereverberation based on deep neural networks
IEEE/ACM Trans. Audio Speech Lang. Proc., vol. 25, pp. 102-111, 2017.

[12] K. Han, et al.
Learning spectral mapping for speech dereverberation and denoising
IEEE/ACM Trans. Audio Speech Lang. Proc., vol. 23, pp. 982-992, 2015.

[13] Simon Leglaive, Umut Simsekli, Antoine Liutkus, Laurent Girin, Radu Horaud
Speech Enhancement with Variational AUto-Encoders and Alpha-stable Distributions
IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), Brighton, UK, May 2019, pp. 541-545

[14] J. Hershey, Z. Chen, J. Le Roux, and S. Watanabe
Deep clustering: Discriminative embeddings for segmentation and separation
in Proceedings of ICASSP, pp. 31-35, 2016.

[15] Dong Yu, Morten Kolbæk, Zheng-Hua Tan and Jesper Jensen
Permutation invariant training of deep models for speaker-independent multi-talker speech separation
2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

[16] Zhuo Chen, Yi Luo, Nima Mesgarani
Deep attractor network for single-microphone speaker separation
2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

[17] Cong Han, Yi Luo, and Nima Mesgarani
Online Deep Attractor Network for Real-time Single-Channel speech separation
2019 IEEE International Conference on Acoustics, Speech and Signal

Processing (ICASSP)

[18] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis
Deep learning for monaural speech separation
in Proceedings of ICASSP, pp. 1581-1585, 2014.

[19] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai
Raw waveform-based speech enhancement by fully convolutional networks
arXiv:1703.02205v3, 2017

[20] Yi Luo, Nima Mesgarani
TasNet: Time-domain Audio Separation for real-time, single-channel speech separation
inforarXiv:1711.00541v1, 2017

[21] Yi Luo, Nima Mesgarani
Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation
arXiv:1809.07454v3, 2019

[22] Deepak Baby and Sarah Verhulst
SERGAN: Speech Ehancement using Relativistic Generative Adversarial Networks with Gradient penalty
2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

[23] Shota Inoue ; Hirokazu Kameoka ; Li Li ; Shogo Seki ; Shoji Makino
Joint Separation and Dereverberation of Reverberant Mixtures with Multichannel Variational Autoencoder
2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

[24] H. Kameoka, L. Li, S. Inoue, and S. Makino
Supervised determined source separation with multichannel variational autoencoder
Neural Comput., vol. 31, no. 9, pp. 1891–1914, 2019.

[25] Mathew Shaji Kavalekalam, Jesper K. Nielsen, Mads G. Christensen and Jesper B. Boldt
Hearing Aid-controlled Beamformer for Binaural Speech Enhancement using a model-based approach
2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

[26] Nico Goßling, Simon Doclo
RTF-Steered Binaural MVDR BF incorporating an external microphone for dynamic acoustic scenarios
2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

[27] Wenqiang Pu1, Jinjun Xiao, Tao Zhang, Zhi-Quan Luo
A Joint Auditory Attentino Decoding and Adaptive Binaural Beamforming Algorithm for Hearing Devices
2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)