Speech emotion recognition with deep learning beamforming on a distant human-robot interaction scenario

Ricardo García[1], Rodrigo Mahu[1], Nicolás Grágeda[1], Alejando Luzanto[1], Nicolás Bohmer[1], Carlos Busso[2], Néstor Becerra Yoma[1]

[1] Speech Processing and Transmission Lab., Electrical Engineering Department, University of Chile

[2] MSP Lab, Department of Electrical and Computer Engineering, The University of Texas at Dallas

nbecerra@ing.uchile.cl

Abstract

Human-robot interaction (HRI) is becoming a truly relevant topic imposing many challenges for state-of-the-art speech technology. This paper describes the first evaluation of speech emotion recognition (SER) technology with non-acted speech data recorded in a real indoor HRI scenario using deep learning-based beamforming technologies. The results presented show that deep learning beamforming gives in average an average concordance correlation coefficient (CCC) that is 15.03% higher than the ordinary minimum variance distortionless response (MVDR) beamformer when the SER system was trained with simulated conditions, which included an acoustic model of the testing HRI environment. Training by simulating the test scenarios and testing with real HRI static data provides on average an average CCC that is just 22.5% smaller than the ideal condition where training and testing were performed with the original MSP-Podcast database. This suggests the possibility to train SER engines with methods that emulates complex testing scenarios without recording further data.

Index Terms: speech emotion recognition; deep learning beamforming; human-computer interaction.

1. Introduction

Social interaction is a very complex challenge for robotics. The difference between human emotional states can be as subtle as "a simple wink, or an upward inflection in a single phoneme" depending on the cultural context [1]. Robotic systems will need to combine multiple input modalities. Nevertheless, some of these inputs, such as physiological signals, require wearable sensors that may be invasive from the user's point of view. In addition, image processing is not always possible depending on the operating conditions. In contrast, speech conveys an enormous amount of linguistic and paralinguistic information (e.g., prosody). Beyond voice commands to robots, speech is a window into the psychological, physical, and emotional state of humans.

Social user profiling is essential for Human-Robot Interaction (HRI): robots are expected to be able to recognize the intentions and goals behind the user's actions to adapt their behavior to them [2]. Within social user profiling, the concept of emotion recognition arises, which seeks to dynamically detect the emotional state of the user during the interaction.

Most of the research in speech emotion recognition (SER) is focused on Human-Computer Interaction (HCI) [3], assuming the user is directly next to the microphone. However, in this case, the influence of the acoustic channel is neglected. Only a few studies have tested distant SER in noisy environments. The most used techniques to address this challenge are the selection of features that are more robust to distance distortions and the creation of encoder-decoder models, which are known to be robust in tasks involving various types of distortions. In [4], 48 low-level descriptors (LLD) were selected and extracted per frame, and passed through a long short-term memory (LSTM) network for final classification. The test environment of this study is a meeting room with seven fixed microphones distributed throughout the room. They performed spectral and temporal filtering. However, no beamforming technique was used. In [5], a metric was employed to determine the distortion of the features according to the distance to the microphone. In addition, they trained their classifier with convoluted audio with artificially generated room impulse responses (RIRs) and used the weighted prediction error (WPE) algorithm to remove reverberation from the test audios and Coherent-to-Diffuse Power Ratio Estimation (CDR) to perform noise cancelling. However, the implementation of the system with a robot was not explored. A feature acquisition technique using a robotic platform with a Kinect mounted was evaluated in Chen et al. [6]. Nevertheless, the test database is acted by volunteers from their own research lab and has only 500 utterances. Furthermore, the study does not address the effects of external noise, which is important to consider since robots, which can generate noise during operation, are crucial in both industrial tasks [7], [8] and butler or personal assistant tasks [6, 9]. Although there is a consensus on the importance of HRI, there are few studies that analyze the effect of this kind of environment on the acoustic channel in systems that use voice as input. In [10], the first evaluation of SER technology with non-acted speech data recorded in a real indoor HRI scenario was presented. The study evaluated the delay-and-sum and MVDR beamforming techniques.

The ability of traditional beamforming approaches to decrease reverberation and noise is limited [11]. Some studies [12][13] compare the application of different beamforming techniques for an ASR system on a robotic platform, achieving improvements with respect to the base cases. Surprisingly, the performance of SER models in complex HRI scenarios has hardly been tested so far except for [10].

In this paper we explore two deep learning-based beamforming techniques (DL-BF) in the context of SER in HRI: Self Attentive MVDR (SA-MVDR) and Self Attentive RNN (SA-RNN). The first method was proposed in [14] and consists of two stages. It is an implementation of MVDR, where in the first stage noise and speech covariances are estimated instantaneously using Ideal Ratio Masks (IRM) [15][16] and Conv-TasNet [17]. In the second step, the instantaneous covariances are used by a transformer in order to estimate the MVDR weights. The second method is SA-RNN, which was

proposed in [18], and it also computes noisy and speech covariance matrices but using Complex Ratio Filters (CRF) [19] with Conv-TasNet. Subsequently, a GRU network [20] and two single attention layers with a final linear layer are used in order to estimate the weights.

In this paper, two state-of-the-art DL-BF techniques, i.e. SA-MVDR and SA-RNN, are evaluated in combination with SER system in a real HRI scenario. To do so, a scheme is proposed to train both the DL-BF methods and the SER engine using acoustic models of the indoor environment. According to the literature, DL-BF schemes outperforms traditional beamforming methods but this comparison has hardly been carried out in SER evaluations.

2. Robotic platform and recording settings

We used the publicly available MSP-Podcast corpus (version 1.9), collected by the Multimodal Signal Processing Laboratory at the University of Texas in Dallas. It has 86,389 speech turns, corresponding to 137 hours of speech annotated with emotional labels. Each speech turn has emotional labels for attribute-based descriptors (valence, activation, and dominance) and categorical labels (happiness, surprise, contempt, neutral, anger, fear, disgust, sadness, and others) that were annotated via crowdsourcing. We employed the testing data recorded in [10] with a static HRI scenario shown in Fig. 1. This data is denoted as *HRI-static*, and was recorded by playing back the MSP-Podcast partition test composed of more than 32 hours of audio. The PR2 robot and a Microsoft Kinect attached to its head (Fig. 2) was employed to record the audios. The robot was placed at P2 (Fig. 1) looking directly towards the speech source (0°) which was 2m away from the robot. The noise sources were located at -45° and 45°. The SNR at P2 was calibrated to be equal to 5dB.

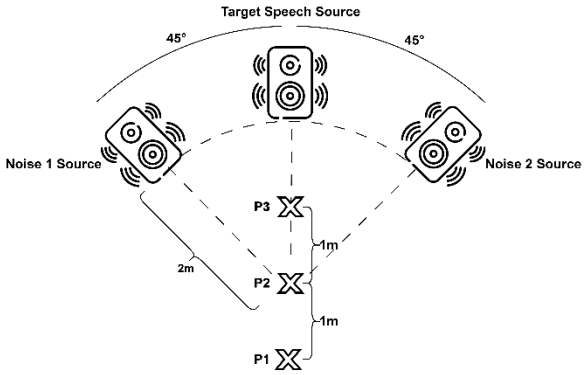

Figure 1: *Diagram of the HRI testbed.*

3. Proposed SER System for HRI with DL-BF schemes

In this paper, we propose the framework shown in Fig. 3. We use acoustic modelling in two stages. The first one is for training the DL-BF schemes, and the second one is for training the SER model. Both acoustic models are not necessarily the same. We assume that the direction of arrival (DOA) can be accurately estimated with computer vision so it can be used by the beamforming methods independently of the acoustic conditions. Finally, it is considered that the indoor acoustic environment can be characterized precisely with additive noise and RIRs measured in the target indoor environment as in [12], and then can be employed by the DL-FB schemes and the SER model.

Figure 2: *Side view of the testbed*

The indoor acoustic model employed here and proposed in [12] makes use of RIRs obtained in static conditions and additive noise that was also played back by loudspeakers. The original training data and additive noise were convoluted with the corresponding RIRs before being artificially added [10]. By doing so, the resulting training dataset would better represent the real indoor HRI conditions.

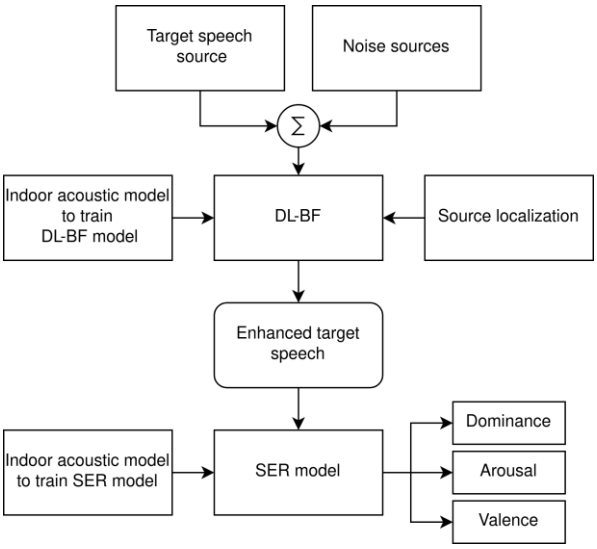

Figure 3: *Proposed SER system with DL-BF schemes.*

3.1. Training of the SER System using Ladder network

The same procedure adopted in [10] was employed to train the Ladder network-based system for SER. The first training dataset was the original MSP-Podcast corpus version 1.9 denoted as Original-training-data. The second training dataset was the MSP-Podcast corpus convoluted with RIRs estimated in the testing indoor HRI environment with artificially added noise as in [10]. We referred to this training dataset as Simulated-data. Three sets of 63 RIRs per each Microsoft Kinect microphone were obtained with the PR2 robot positioned at P1, P2, and P3 (Fig. 1) and by orienting the robot head at 21 different angles with respect to the source. The head angle was varied from -50° to 50° in 5° steps. As mentioned above, the 0° angle corresponds to the PR2 robot head looking directly toward the speech source. The RIRs were computed

with the swept-sine method proposed in Farina [21]. An exponential sweep from 64 Hz to 8 kHz sine functions was generated and played back with a studio loudspeaker located at the Target Speech, Noise 1, and Noise 2 source positions (see Fig. 1). The audio of the reproduced sweep was recorded with the four Kinect microphones and an impulse response was estimated for each channel. The three sets of 63 RIRs were named according to where the studio loudspeaker was positioned to reproduce the swept sine functions: *RIR-Target Source*, *RIR-Noise1 Source* and *RIR-Noise2 Source*.

The Ladder network was trained with multitask learning, jointly predicting arousal, valence, and dominance as in [22]. The input to the network is the ComParE feature set [23], which has 6,373 high-level descriptors (HLD), regardless of the audio duration of the speech segment. For training, 100 epochs were run with learning rate equal to 0.0001 on an NVDIA 3080 GPU. We chose the best one of them based on the validation subset CCC. Five training conditions were evaluated: *Original training-data*; *Simulated-data*; *Simulated-data* combined with *MVDR*; *Simulated-data* combined with *SA-MVDR*; and *Simulated-data* combined with *SA-RNN*. In each condition, 10 training iterations were carried out, and the averages of the results obtained are presented.

3.2. DL-BF training

The DL-BF schemes employed the Aurora-4 database [24], which contains 7,138 (train set), 330 (validation set), and 330 (test set) recordings. For noise addition, we employed the DEMAND database [25], which contains 18 kinds of noise, 14 were used for training, two were employed for validation, and two were for testing. Then, the noise produced by PR2 was added [25]. The proportion used to create the dataset is 25% of the total utterances with Noise source 1 only, 25% with Noise source 2 only, and the remaining 50% a mixture of both sources with a ratio between -5dB and 5dB. The speech samples, Noise source 1 and Noise source 2 were convoluted with *RIR-Target_Source*, *RIR-Noise1_Source* and *RIR-Noise2_Source*, respectively, before being added. The convolved noise sources with the corresponding RIRs were added to the robot noise at a ratio between 5dB to 10dB. Finally, the resulting noise was added to the speech signal convolved with the corresponding RIRs with an SNR between 0dB and 10dB. With this methodology, a total of 14,276 (train set), 1,980 (validation set) and 1,320 (test set) utterances were generated. We emploed the scale-dependent signal-to-noise ratio loss function [26] for training: $Loss = -10log\left(\frac{||s||^2}{||s-\hat{s}||^2}\right)$. Both the CRF estimator and the transformer were trained with a batch size of 8 and the ADAM optimizer for SA-MVDR. The learning rate was made equal to 10^{-4} and $5x10^{-5}$ for the first estimator and second estimator, respectively. The training of the SA-RNN beamformer used a batch size of 1, the ADAM optimizer with a learning rate of 10^{-4} and a gradient clipping with max norm of 10.

3.3. Testing databases

Three testing data were employed. The first one is composed of speech samples from the test partition of the MSP-Podcast corpus and is denoted as *Original-testing-data*. The second test condition corresponds to *Simulated-data* generated with the same procedure as the corresponding training data (section 3.1). Finally, the third dataset corresponds to speech samples recorded with the real HRI platform in static condition, *HRI-Static* (section 2).

3.4. About the baseline system

The performance of the baseline system is competitive with those published elsewhere. MSP-Podcast corpus is a naturalistic database with several sentences with ambiguous emotional content. Therefore, it is not straightforward to compare the results with the performance observed in other databases with more controlled settings (e.g., acted recordings). The Ladder networks were obtained by considering a state-of the-art approach with the same code used in [22]. Also, the newest versions of the MSP-Podcast corpus have incorporated more challenging samples, making the test set more difficult. They include more speakers in the test set, increasing the variability in the set. For example, Lin et al. [27], using version 1.10, recently reported similar values to the ones we are reporting in this study.

3.5. Original training data & real HRI testing

Table 1 shows the concordance correlation coefficient (CCC) and SNR when the Ladder network was trained with *Original training-data* and tested with *HRI-static*. Additionally, Table 1 also shows the results when *HRI-static* was also processed with MVDR, SA-RNN and SA-MVDR. As can be seen in Table 1, testing with the *Original testing-data* led to averaged CCC and SNR equal to 0.416 and 14.34dB, respectively. They are the optimal values that can be obtained according to the framework adopted here. When the testing data corresponds to *HRI-static*, we observed the greatest degradations in the average CCC and SNR with reductions of 74% and 8.84dB, respectively. When applied to the test data only, beamforming technology increases both the average CCC and SNR when *HRI-static* is employed. Ordinary MVDR, SA-RNN and SA-MVDR led to increases in the average CCC of 127%, 132% and 136%, respectively (statistically significant with $p<10^{-6}$). Regarding SNR, ordinary MVDR, SA-RNN and SA-MVDR led to increases of 4.01dB, 4.89dB and 3.85db, respectively.

Table 1: *Results obtained when the Ladder network was trained with Original-training-data*.

Train type	Test type	SNR	Aro	Dom	Val
Original testing	***Original testing***	14.34	0.571	0.461	0.216
HRI-Static	***HRI-Static***	5.46	0.172	0.112	0.042
HRI-static + MVDR	***HRI-static + MVDR***	9.47	0.358	0.305	0.078
HRI-static + SA-RNN	***HRI-static + SA-RNN***	14.36	0.367	0.311	0.081
HRI-static + SA-MVDR	***HRI-static + SA-MVDR***	9.31	0.377	0.334	0.061

Observe that the highest improvements in average CCC were achieved with SA-RNN and SA-MVDR even though the

former provides a much higher increase in SNR that the latter, which in turn suggests that the artifact introduced by the beamforming schemes is not necessarily correlated to the SNR improvement. It is worth highlighting that SA-RNN and SA-MVDR schemes were trained with the same database, under the same conditions, but SA-RNN was trained on an end-to-end basis since the reconstruction of the audios in training takes place by using the output of the ConvTasnet model as input. In contrast, SA-MVDR reconstructs the audios in training with oracle covariance matrices. Moreover, ordinary MVDR provides a higher SNR than SA-MVDR but a lower average CCC. This result must be due to the fact that SA-MVDR is trained on a close loop basis to reconstruct the reference signal. Finally, it can be mentioned that the DL-BF methods present an average improvement 3.43% with respect to MVDR (statistically significant with $p < 10^{-2}$).

3.6. Models trained & tested with simulated data

Table 2 shows the results when the Ladder network was trained and tested with *Simulated-data* according to sections 3.1 and 3.3. Beamforming methods were also included in training and testing. As can be seen in Table 2, training and testing with *Simulated-data+MVDR*, *Simulated-data+SA-RNN* and *Simulated-data+SA-MVDR* decreases the difference in average CCC with respect to *Original training-data/Original testing-data* when compared to Table 1 (Fig. 5). In average, training and testing with *Simulated-data+MVDR, Simulated-data+SA-RNN* and *Simulated-data+SA-MVDR* provided an average increase in average CCC equal to 30.56% when compared to testing with *HRI-static+MVDR, HRI-static+SA-MVDR*, and *HRI-static+SA-MVDR* and training with *Original-training-data* (Fig. 5). When compared to the baseline result achieved with *Original-training-data/Original-testing-data*, training and testing with *Simulated-data+MVDR, Simulated-data+SA-RNN* and *Simulated-data+SA-MVDR* provides an average reduction in average CCC as small as 20.8% (Fig. 5). This difference must be due to the distortion introduced by non-controlled variations of reverberation and additive noise that are not fully represented by the acoustic model employed here.

Table 2: *Results obtained when the Ladder network was trained and evaluated with simulated conditions.*

Train type	Test type	Aro	Dom	Val
Simulated + MVDR	***Simulated + MVDR***	0.482	0.350	0.101
Simulated + SA-RNN	***Simulated + SA-RNN***	0.515	0.375	0.130
Simulated + SA-MVDR	***Simulated + SA-MVDR***	0.500	0.374	0.138

3.7. Models trained with simulated & tested in real HRI

As can be seen in Table 3 and Fig. 5, training with simulated conditions and testing with real HRI static data provides practically the same results as training and testing with simulated conditions. Particularly with SA-RNN and SA-MVDR, there is not a significant difference between both scenarios. It is important to mention that the average improvement in average CCC obtained by training with simulated condition and testing with *HRI-static+MVDR, HRI-static+SA-RNN*, and *HRI-static+SA-MVDR* (Table 3) compared to training with *Original training-data* and testing with the same conditions (Table 1) is on average 27.7% (statistically significant with $p < 10^{-6}$).

Table 3: *Results obtained when training with simulated conditions and rested with real HRI data.*

Train type	Test type	Aro	Dom	Val
Simulated + MVDR	***HRI-Static + MVDR***	0.440	0.341	0.099
Simulated +SA-RNN	***HRI-Static + SA-RNN***	0.492	0.370	0.126
Simulated + SA-MVDR	***HRI-Static + SA-MVDR***	0.494	0.386	0.122

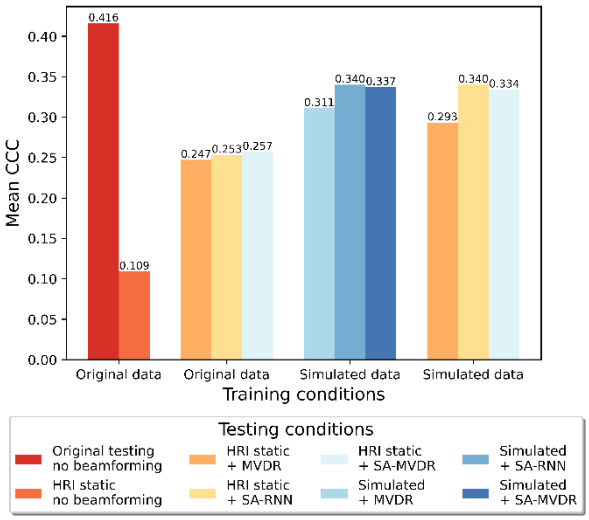

Figure 4: *Averages CCC of the Ladder Network evaluation with the different training and test conditions.*

4. Conclusions

This paper describes the first evaluation of SER technology with non-acted speech data recorded in a real indoor HRI scenario using deep learning-based beamforming technology, i.e. SA-RNN and SA-MVDR. The results presented here show that SA-RNN and SA-MVDR provides in average a CCC that is 15.0% higher than the ordinary MVDR beamformer when the SER system was trained with simulated conditions (statistically significant with $p<10^{-3}$), which included an acoustic model of the testing HRI indoor environment and the response of the beamforming technologies. Training by simulating the test scenarios and testing with real HRI static data, MVDR, SA-RNN and SA-MVDR provide in average an average CCC that is just 22.5% smaller than the ideal condition where training and testing were performed with the original MSP-Podcast database (statistically significant with $p<10^{-6}$). This result is interesting because unveil the possibility to train SER engines with methods that emulates complex testing scenarios without the need for recording further data. To show the difficulty of the task addressed here it should be noted that, when training with the original MSP-Podcast data, testing with real HRI static data provides a reduction in average CCC of 74% when compared to testing with the original MSP-Postcast data. Finally, addressing more complex HRI scenarios is proposed as future research

5. Acknowledgements

This research was funded by ANID/FONDECYT (Chile) grant No. 1211946.

6. References

[1] G. Z. Yang *et al.*, 'The grand challenges of science robotics', *Science Robotics*, vol. 3, no. 14. 2018. doi: 10.1126/scirobotics.aar7650.

[2] S. Rossi, F. Ferland, and A. Tapus, 'User profiling and behavioral adaptation for HRI: A survey', *Pattern Recognit Lett*, vol. 99, 2017, doi: 10.1016/j.patrec.2017.06.002.

[3] M. Shah Fahad, A. Ranjan, J. Yadav, and A. Deepak, 'A survey of speech emotion recognition in natural environment', *Digital Signal Processing: A Review Journal*, vol. 110. 2021. doi: 10.1016/j.dsp.2020.102951.

[4] A. Salekin *et al.*, 'Distant Emotion Recognition', *Proc ACM Interact Mob Wearable Ubiquitous Technol*, vol. 1, no. 3, 2017, doi: 10.1145/3130961.

[5] M. Y. Ahmed, Z. Chen, E. Fass, and J. Stankovic, 'Real time distant speech emotion recognition in indoor environments', in *ACM International Conference Proceeding Series*, 2017. doi: 10.1145/3144457.3144503.

[6] L. Chen, W. Su, Y. Feng, M. Wu, J. She, and K. Hirota, 'Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction', *Inf Sci (N Y)*, vol. 509, 2020, doi: 10.1016/j.ins.2019.09.005.

[7] J. Berg, A. Lottermoser, C. Richter, and G. Reinhart, 'Human-Robot-Interaction for mobile industrial robot teams', in *Procedia CIRP*, 2019, vol. 79. doi: 10.1016/j.procir.2019.02.080.

[8] N. Kousi, C. Stoubos, C. Gkournelos, G. Michalos, and S. Makris, 'Enabling human robot interaction in flexible robotic assembly lines: An augmented reality based software suite', in *Procedia CIRP*, 2019, vol. 81. doi: 10.1016/j.procir.2019.04.328.

[9] J. Miseikis *et al.*, 'Lio-A Personal Robot Assistant for Human-Robot Interaction and Care Applications', *IEEE Robot Autom Lett*, vol. 5, no. 4, 2020, doi: 10.1109/LRA.2020.3007462.

[10] N. Grageda, E. Alvarado, R. Mahu, C. Busso, and N. Yoma, "Distant speech emotion recognition in an indoor human-robot interaction scenario," 08 2023, pp. 3657–3661.

[11] K. U. Simmer, J. Bitzer, and C. Marro, 'Post-Filtering Techniques', 2001. doi: 10.1007/978-3-662-04619-7_3.

[12] J. Novoa, R. Mahu, J. Wuth, J. P. Escudero, J. Fredes, and N. B. Yoma, 'Automatic Speech Recognition for Indoor HRI Scenarios', *ACM Trans Hum Robot Interact*, vol. 10, no. 2, 2021, doi: 10.1145/3442629.

[13] A. Díaz, R. Mahu, J. Novoa, J. Wuth, J. Datta, and N. B. Yoma, 'Assessing the effect of visual servoing on the performance of linear microphone arrays in moving human-robot interaction scenarios', *Comput Speech Lang*, vol. 65, 2021, doi: 10.1016/j.csl.2020.101136.

[14] T. Ochiai, M. Delcroix, T. Nakatani, and S. Araki. 2023. Mask-Based Neural Beamforming for Moving Speakers With Self-Attention-Based Tracking. *IEEE/ACM transactions on audio, speech, and language processing* 31, (2023), 835–848.

[15] S. Balasubramanian, R. Rajavel, and A. Kar. 2023. Ideal ratio mask estimation based on cochleagram for audio-visual monaural speech enhancement. *Applied Acoustics* 211, (2023), 109524.

[16] Z. Xu, S. Elshamy, Z. Zhao, and T. Fingscheidt. 2021. Components loss for neural networks in mask-based speech enhancement. *EURASIP Journal on Audio, Speech, and Music Processing* 2021, 1 (2021), 1–20.

[17] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time - frequency magnitude masking for speech separation," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. PP, pp. 1–1, 05 2019.

[18] X. Li, Y. Xu, M. Yu, S.-X. Zhang, J. Xu, B. Xu, and D. Yu, "MIMO Self-Attentive RNN Beamformer for Multi-Speaker Speech Separation," in Proc. Interspeech 2021, 2021, pp. 1119–1123.

[19] W. Mack and E. Habets, "Deep filtering: Signal extraction and reconstruction using complex time-frequency filters," IEEE Signal Processing Letters, vol. PP, pp. 1–1, 11 2019.

[20] K. Cho, B. Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," 09 2014.

[21] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," 11 2000.

[22] S. Parthasarathy and C. Busso, "Semi-supervised speech emotion recognition with ladder networks," IEEE/ACM Trans. Audio, Speech and Lang. Proc., vol. 28, p. 2697–2709, oct 2020. [Online]. Available: https://doi.org/10.1109/TASLP.2020.3023632

[23] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," 08 2013, pp. 148–152.

[24] D. Pearce and H.-G. Hirsch, "The aurora experimental framework for the performance evaluations of speech recognition systems under noisy condition," vol. 4, 10 2000, pp. 29–32.

[25] J. Thiemann, N. Itoy E. Vincent, «DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments». Zenodo, jun. 09, 2013. doi: 10.5281/zenodo.1227121.

[26] J. Le Roux, S. Wisdom, H. Erdogan, and J. Hershey, "Sdr – half-baked or well done?" 05 2019, pp. 626–630.S. W.-C.

[27] Lin and C. Busso, "Role of lexical boundary information in chunk-level segmentation for speech emotion recognition," in ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5.