

Distant speech emotion recognition in an indoor human-robot interaction scenario

Nicolás Grágeda¹, Carlos Busso², Eduardo Alvarado¹, Rodrigo Mahu¹, Néstor Becerra Yoma¹

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

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

nbecerra@ing.uchile.cl

Abstract

Social robotics and human-robot partnership are becoming very relevant topics defining many challenges for state-of-the-art speech technology. This paper presents the first evaluation of speech emotion recognition (SER) technology with non-acted speech data recorded in a real indoor human-robot interaction (HRI) scenario. The challenge is typified by distant speech processing, reverberation, and additive external and robot engine noise. We train and evaluate a machine learning-based model based on simulated acoustic modelling that includes room impulse responses (RIRs), external noise, and beamforming response. We observe increased performance in the prediction of arousal, valence, and dominance with the proposed training procedure combined with delay-and-sum and minimum variance distortionless response (MVDR), with gain as high as 180%, compared with the result obtained with the model trained with the original data in controlled environments. Moreover, the degradation achieved when compared with the original matched training/testing condition is just 39%.

Index Terms: speech emotion recognition, human-computer interaction.

1. Introduction

Social interaction is a very complex challenge for robotics, in part because it requires effectively recognizing or detecting gaze directions, facial expressions, linguistic content, and prosody of speech, and then acting accordingly. Depending on the cultural context, the difference between human emotional states can be as subtle as "a simple wink, or an upward inflection in a single phoneme." [1]. To achieve this purpose, robotic systems will need to combine multiple input modalities. However, 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), because the robots are expected to be able to recognize the intentions and goals behind the user's actions, in order to adapt their behavior to them [2]. In addition, social profiling also refers to the ability to recognize social phenomena, such as commitment, conflict, empathy, interest and emotions, which cannot be observed directly, but must be inferred by examining indirect indicators. Some of these indirect indicators can be

body posture [3], facial expressions [4], gaze direction [5], and voice volume. Within social user profiling, the concept of emotion recognition arises, which seeks to dynamically detect the emotional state of the user during the interaction, because, while a person's emotional profile does not change during a single interaction with the robot, the user may exhibit multiple emotions during the interaction. This continuous detection approach allows the user's profile to be frequently updated.

The process of identifying human emotions using the voice, mainly non-verbal elements of the voice is defined as speech emotion recognition (SER). The vast majority of the research in this discipline is focused on Human-Computer Interaction (HCI) [6], 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. Salekin et al. [7] selected 48 low-level descriptors (LLD), which were 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. Ahmed et al. [8] employed a metric 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 use 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. [9]. Nevertheless, the test database is acted upon 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 noisy during operation, are crucial in both industrial tasks [10], [11] and butler or personal assistant tasks [9], [12]. 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.

One of the most popular architectures in SER is the ladder network, especially those using semi-supervised training [13–16]. This type of network consists of an encoder-decoder scheme with lateral connections between these two modules. The encoder is trained to perform the classification or linear regression task (as the case may be) with an input to which

noise, usually Gaussian, is added at each of the layers. While the decoder is trained to perform a reconstruction of the original input (before adding the noise) of each layer. Leem et al. [17] tested a ladder network implementation in a noisy environment using the microphone of a smartphone, a speaker that reproduces speech, and another speaker at an opposite end that reproduces noise. The setting was fixed, as the microphones and speakers did not move during the data collection.

This paper addresses the challenging problem of SER using distant speech in the context of HRI. We envision a human-robot collaborative scenario. This scenario also considers the noise in the audio produced by environmental conditions and the robot. The proposed approach evaluates beamforming techniques combined with source localization to deal with distant speech. We address the generalization of the SER model to a new domain by using the semi-supervised strategy based on the ladder network. The approach is implemented and evaluated with the MSP-Podcast corpus, which is the "largest naturalistic emotional dataset in the community" [18]. This database, unlike most databases [19-21], contains fragments of non-acted audio, in normal speech environments, so it better matches real world recordings. It is impossible to fully describe the complexity of human emotions using a few categorical labels [22][23], which is why emotion recognition in a continuous three-dimensional space of emotional attributes (arousal, dominance, and valence) is selected as the task for this research.

Beamforming is one of the spatial filtering techniques used successfully to enhance signals coming from a certain direction relative to a set of microphones, reducing noise and interference coming from other directions. However, the ability of traditional beamforming approaches to decrease reverberation and diffuse noise is limited [24]. Some studies [25][26] compare the application of different beamforming techniques for an ASR system on a robotic platform, achieving improvements with respect to the base cases. This paper evaluates two widely employed beamforming techniques with SER in a complex, non-stationary HRI scenario. These techniques are the well-known delay-and-sum (D&S) scheme [27], and the minimum variance distortionless response (MVDR) method [28]. Surprisingly, the performance of SER models in complex HRI scenarios has hardly been tested so far. However, there are studies on the effect of a complex scenario in HRI for speech-to-text task [25]. Based on these studies, we propose a setup for re-recording the test partition of the MSP-Podcast database. The proposed testbed illustrates the generic problem of HRI in mobile robotics regarding SER including distant speech processing, external noise sources, and noise coming from the engine of the robot. In addition, we simulate target source localization for beamforming, which in turn is feasible with the sensors mounted on the robot (e.g., cameras), to steer the main mic array lobe. This method is a first step towards more complete integration of SER to complex HRI scenarios. This paper addresses the acoustic channel modelling problem by using the RIR to simulate a real environment in the training database.

The main contribution of this study is the proposed setting to simulate scenarios for HRI during social, collaborative interactions and the evaluation of state-of-the-art techniques for noise robustness and semi-supervised SER solutions to address this challenging problem. An important contribution is also the database re-collected using the proposed setting, which will be shared with the community.

2. Proposed framework

In HRI situations, robots can use sensors such as cameras to determine the position of the target speaker and, therefore, have a more precise estimate of the angle of incidence or direction of arrival (DOA) corresponding to the speech source [26]. By doing so, it is possible to avoid the error introduced by reverberation in indoor scenarios.

2.1. Proposed system

This paper proposes the framework in Fig. 1 to address the problem of SER in mobile HRI to cope with the challenges imposed by the source-microphone distance, noise sources, and time-varying acoustic channel (TVAC) [25]. The following assumptions are included in this framework: first, the angular position of the target source can be estimated accurately independently of the error introduced by indoor reverberation; second, beamforming technology can use the target speaker's angular position to deliver improved spatial filtering; third, TVAC in an indoor environment can be addressed by making use of RIRs obtained in static conditions as in Novoa et al. [25].

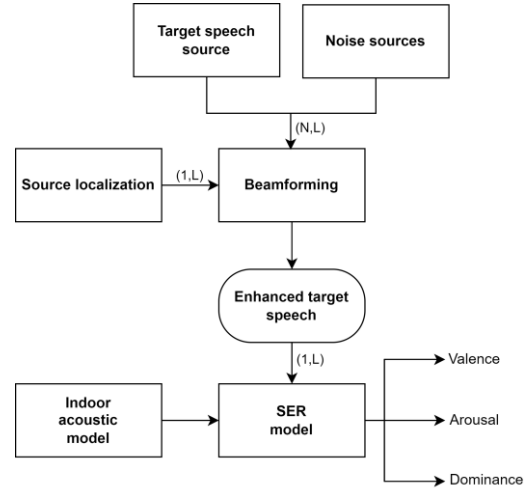


Figure 1: Proposed SER system in HRI. Where N is the number of channels and L is the length of the audio.

Two beamforming techniques are considered in this study: D&S and MVDR. In the case of MVDR, the noise covariance matrix in speech segments was made equal to the interpolation of the matrices corresponding to the pre and post noise intervals. For the purpose of this paper, indoor acoustic modelling (AM) represents the reflections of both the target speech and the additive external noise signals using RIRs experimentally obtained in the same environment as in the HRI test datasets.

As indicated in Fig. 1, to improve the performance of SER models in real HRI indoor scenarios, the indoor AM is modelled similarly to Novoa et al. [25] with RIRs obtained in static conditions and additive noise. The original training data and additive noise are convoluted with the corresponding RIRs before being artificially added. The resulting training dataset represents better real HRI conditions.

2.2. Robotic platform and recording settings

We use 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.

The test partition of the corpus was played back in complex real HRI scenarios. This test partition has 21,560 turns of speech and accumulates more than 32 hours of audio. The HRI testbed was implemented with the PR2 robot equipped with a Microsoft Xbox 360 Kinect sensor mounted on top of its head. As shown in Fig. 4, we use one speech and two noise sources, each one located 2m away from point $P2$. The noise sources are 45° on either side of the speech source. The average recording signal-to-noise ratio (SNR) was adjusted to be equal to 5dB measured at point $P2$. For the static scenario, the PR2 robot stays still at $P2$, with its head pointing directly to the speech source.

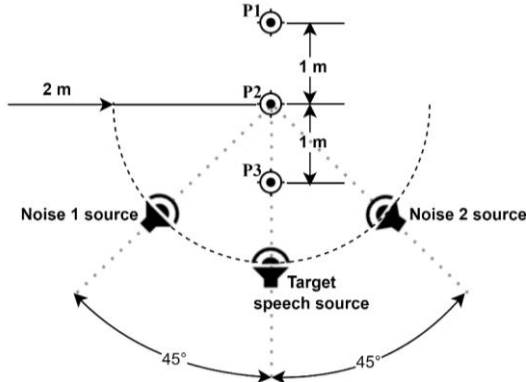


Figure 2: Diagram of the testbed



Figure 3: Side view of the testbed

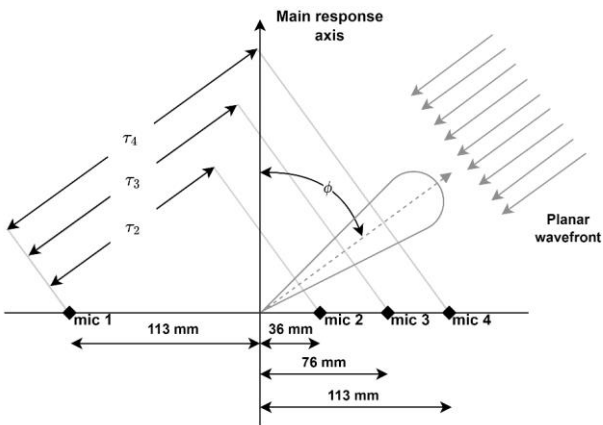


Figure 4: Microphone array geometry of the Microsoft Kinect, where: τ_n is the time delay between microphone n and microphone 1; and, ϕ is the look direction or DOA.

In contrast to Novoa et al. [25], three sets of 63 RIR per each Microsoft Kinect microphone were obtained with the PR2 robot positioned at $P1$, $P2$, and $P3$ (Fig. 2) 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. 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 [29]. An exponential sweep from 64 Hz to 8 kHz sine functions was generated and played back with a studio loudspeaker located at the points *target*, *Noise 1* and *Noise 2* source positions (see Fig. 2). The audio of the reproduced sweep was recorded with the four Microsoft Kinect microphones. An impulse response was estimated for each channel by convoluting the corresponding recorded signal with the time reversal of the original exponential sinusoidal sweep. 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*.

3. Experiment and results

3.1. Training datasets

The SER architectures evaluated here were trained with two types of data. First, we use the original MSP-Podcast corpus, which we referred to as *Original_training*. The second training data corresponds to the same audios but convoluted with the RIRs estimated as aforementioned and with noise added artificially to emulate the real HRI testing scenario. We referred this setting to as *Simulated_training*.

A simulated training dataset was generated with the set *RIR-Target_Source* of impulse responses as follows: 25% of the data from each partition was convoluted with the RIR obtained at $P1$ while the robot head looks directly to the target source. The remaining 75% of the audio files from each partition were convoluted with the remaining 62 RIRs, so that each of these RIRs was used in the same number of simulated audios. Then, the noise was added artificially to the resulting audios at SNRs that were randomly chosen between 10dB and 20dB. The additive noise was obtained as follows: noise segments from DEMAND [30] were convoluted with the impulse responses from *RIR-Noise1_Source* and *RIR-Noise2_Source*; then, they were added with the same ratio and considering the same robot position of the speech signal that they were adding to; and, the resulting external additive noise was summed to the PR2 engine noise at SNRs between -5dB and 5db. Moreover, the resulting reverberated noisy data from the four Microsoft Kinect microphones were delayed and combined with the D&S and MVDR beamforming methods.

3.2. Training of the SER System using Ladder network

We employ the SER architecture based on the ladder network proposed by Parthasarathy and Busso [15]. The network is trained with multitask learning, jointly predicting arousal, valence, and dominance. The input to the network is the ComParE feature set [31], 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 NVIDIA 3080 GPU. We chose the best one of them based on the CCC validation subset.

3.3. Testing databases

We report results with three testing conditions: *Original_testing*, corresponding to audios from the test partition of the MSP-Podcast corpus; *Simulated_testing*, corresponding to audios from the test partition of the MSP-Podcast corpus, which were similarly processed to those in the *Simulated_training*; and *HRI_static*, corresponding to testing audios from the MSP-Podcast corpus re-recorded in the robotic platform in static conditions (see section 3). We assess the use of the beamforming schemes D&S and MVDR with *Simulated_testing* and *HRI_static*.

3.4. Original training data & real HRI testing

Table 1 shows the concordance correlation coefficient (CCC) obtained when Ladder Network was trained with *Original_training* and tested with dynamic testing scenarios. The testing subsets corresponded to *Original_testing*, *HRI_static*, *HRI_static+D&S* and *HRI_static+MVDR*. According to Table 1, the highest degradation in CCC Arousal, CCC Dominance and CCC Valence when compared with *Original_testing* was observed with *HRI_static* with Ladder Network. Beamforming schemes D&S and MVDR increase the SNR and decrease the degradation in CCC for arousal, dominance and valence when compared with the *Original_testing* using Ladder Network. The increase in SNR is equal to 52.75% and 71.25% with D&S and MVDR, respectively. When compared with the *HRI_static*, D&S and MVDR led to an increase in the summation of CCC's equal to 108.77% and 117.09%, respectively, when using Ladder Network.

Table 1: Results obtained with models trained with original data.

Test type	SNR	Aro	Dom	Val	Sum
<i>Original_testing</i>	-	0.629	0.536	0.266	1.431
<i>HRI_static</i>	5.46	0.175	0.066	0.073	0.314
<i>HRI_static + D&S</i>	8.34	0.343	0.233	0.079	0.655
<i>HRI_static + MVDR</i>	9.35	0.313	0.251	0.118	0.681

3.5. Models trained & tested with simulated data

Table 2 shows the results when Ladder Network was trained and tested with the simulated database described in section 3.1 for the real static database. Two training/testing conditions were employed: *Simulated+D&S* and *Simulated+MVDR*, where D&S and MVDR were applied after, as explained above, respectively. Results with *Simulated+D&S* and *Simulated+MVDR* correspond to static conditions and can be compared with results obtained with *Original_training* and *Original_testing*. According to Tables 1 and 2, *Simulated+D&S* and *Simulated+MVDR* with Ladder Network still led to reductions in the summation of the CCC scores equal to 29.15% and 34.44%, respectively, when compared with the *Original_training* and *Original_testing*. Although the conditions in Table 2 are somehow matched, this result suggests that the added noise and reverberation still introduce some uncertainty. Nevertheless, the achieved sums in CCC metrics

are 54.80% and 37.75% greater than those with the testing subsets *HRI_static+D&S* and *HRI_static+MVDR*, which in turn is also caused by the fact that training and testing data were generated in similar conditions.

Table 2: Results obtained with models trained and tested with simulated data.

Train and test type	Aro	Dom	Val	Sum
<i>Simulated+D&S</i>	0.520	0.374	0.120	1.014
<i>Simulated+MVDR</i>	0.492	0.352	0.095	0.939

3.6. Models trained with simulated & tested in real HRI

Table 3 presents the results when Ladder Network was trained with simulated data and tested with real static HRI data. As can be seen in Tables 2 and 3, the difference between the sum of CCC scores obtained with *HRI_static+D&S* and *Simulated_testing+D&S* when Ladder Network was trained with *Simulated_training+D&S* was just 17.65%. A similar result was observed with *HRI_static+MVDR* and *Simulated_testing+MVDR* when the difference in the sum of the CCC metrics was only 6.32%.

Results in Table 3 basically suggest that the simulated training conditions proposed here, represented by subsets *Simulated_training+D&S* and *Simulated_training+MVDR*, are quite close approximations to real static HRI scenarios.

Table 3: Results obtained with models trained with simulated data and tested in HRI static position.

Train type	Test type	Aro	Dom	Val	Sum
<i>Simulated + D&S</i>	<i>HRI_static + D&S</i>	0.426	0.317	0.093	0.836
<i>Simulated + MVDR</i>	<i>HRI_static + MVDR</i>	0.437	0.342	0.100	0.879

4. Conclusions

This paper describes the first evaluation of SER technology with non-acted speech data recorded in a real indoor HRI scenario. The challenge is characterized by distant speech processing, reverberation, and additive external and robot engine noise. We evaluate machine learning training based on simulated acoustic modelling that includes RIRs, external noise and beamforming response. The average increase in the sum of CCC metrics with the proposed training procedure combined with delay-and-sum and MVDR when compared with the result obtained with the model trained with the original data in controlled environments is 166% and 180%, respectively. The degradation obtained when compared with the original matched training/testing condition is just 39%. We propose as future research to test with dynamic real HRI scenarios and other SER classifiers.

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] A. Gaschler, S. Jentzsch, M. Giuliani, K. Huth, J. de Ruyter, and A. Knoll, ‘Social behavior recognition using body posture and head pose for human-robot interaction’, in *IEEE International Conference on Intelligent Robots and Systems*, 2012. doi: 10.1109/IROS.2012.6385460.
- [4] D. R. Faria, M. Vieira, F. C. C. Faria, and C. Premebida, ‘Affective facial expressions recognition for human-robot interaction’, in *RO-MAN 2017 - 26th IEEE International Symposium on Robot and Human Interactive Communication*, 2017, vol. 2017-January. doi: 10.1109/ROMAN.2017.8172395.
- [5] P. Chakraborty, S. Ahmed, M. A. Yousuf, A. Azad, S. A. Alyami, and M. A. Moni, ‘A Human-Robot Interaction System Calculating Visual Focus of Human’s Attention Level’, *IEEE Access*, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3091642.
- [6] 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.
- [7] A. Salekin *et al.*, ‘Distant Emotion Recognition’, *Proc ACM Interact Mob Wearable Ubiquitous Technol*, vol. 1, no. 3, 2017, doi: 10.1145/3130961.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] J. Huang, Y. Li, J. Tao, Z. Lian, M. Niu, and J. Yi, ‘Speech Emotion Recognition Using Semi-supervised Learning with Ladder Networks’, in *2018 1st Asian Conference on Affective Computing and Intelligent Interaction, ACII Asia 2018*, 2018. doi: 10.1109/ACIIAsia.2018.8470363.
- [14] S. Parthasarathy and C. Busso, ‘Ladder networks for emotion recognition: Using unsupervised auxiliary tasks to improve predictions of emotional attributes’, in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2018, vol. 2018-September. doi: 10.21437/Interspeech.2018-1391.
- [15] S. Parthasarathy and C. Busso, ‘Semi-Supervised Speech Emotion Recognition with Ladder Networks’, *IEEE/ACM Trans Audio Speech Lang Process*, vol. 28, 2020, doi: 10.1109/TASLP.2020.3023632.
- [16] J. H. Tao, J. Huang, Y. Li, Z. Lian, and M. Y. Niu, ‘Semi-supervised Ladder Networks for Speech Emotion Recognition’, *International Journal of Automation and Computing*, vol. 16, no. 4, 2019, doi: 10.1007/s11633-019-1175-x.
- [17] S. G. Leem, D. Fulford, J. P. Onnela, D. Gard, and C. Busso, ‘Separation of emotional and reconstruction embeddings on ladder network to improve speech emotion recognition robustness in noisy conditions’, in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2021, vol. 1. doi: 10.21437/Interspeech.2021-1438.
- [18] R. Lotfian and C. Busso, ‘Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech from Existing Podcast Recordings’, *IEEE Trans Affect Comput*, vol. 10, no. 4, 2019, doi: 10.1109/TAFFC.2017.2736999.
- [19] C. Busso *et al.*, ‘IEMOCAP: Interactive emotional dyadic motion capture database’, *Lang Resour Eval*, vol. 42, no. 4, 2008, doi: 10.1007/s10579-008-9076-6.
- [20] A. Metallinou, Z. Yang, C. Chun Lee, C. Busso, S. Carnicke, and S. Narayanan, ‘The USC CreativeIT database of multimodal dyadic interactions: from speech and full body motion capture to continuous emotional annotations’, *Lang Resour Eval*, vol. 50, no. 3, 2016, doi: 10.1007/s10579-015-9300-0.
- [21] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, ‘CREMA-D: Crowd-sourced emotional multimodal actors dataset’, *IEEE Trans Affect Comput*, vol. 5, no. 4, 2014, doi: 10.1109/TAFFC.2014.2336244.
- [22] L. Devillers, L. Vidrascu, and L. Lamel, ‘Challenges in real-life emotion annotation and machine learning based detection’, *Neural Networks*, vol. 18, no. 4, 2005, doi: 10.1016/j.neunet.2005.03.007.
- [23] E. Mower *et al.*, ‘Interpreting ambiguous emotional expressions’, in *Proceedings - 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII 2009*, 2009. doi: 10.1109/ACII.2009.5349500.
- [24] K. U. Simmer, J. Bitzer, and C. Marro, ‘Post-Filtering Techniques’, 2001. doi: 10.1007/978-3-662-04619-7_3.
- [25] 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.
- [26] 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.
- [27] M. Omologo, M. Matassoni, and P. Svaizer, ‘Speech Recognition with Microphone Arrays’, in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 331–353. doi: 10.1007/978-3-662-04619-7_15.
- [28] J. Bitzer and K. U. Simmer, ‘Superdirective Microphone Arrays’, in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 19–38. doi: 10.1007/978-3-662-04619-7_2.
- [29] A. Farina, ‘Simultaneous measurement of impulse response and distortion with a swept-sine technique’, *Proc. AES 108th conv, Paris, France*, no. 1, 2000.
- [30] J. Thiemann, N. Ito, and E. Vincent, ‘The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings’, in *Proceedings of Meetings on Acoustics ICA2013*, 2013, vol. 19, no. 1, p. 35081.
- [31] B. Schuller *et al.*, ‘The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism’, in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2013. doi: 10.21437/interspeech.2013-56.