Computer Speech and Language 89 (2025) 101666

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Speech emotion recognition in real static and dynamic human-robot interaction scenarios

Nicolás Grágeda [a], Carlos Busso [b], Eduardo Alvarado [a], Ricardo García [a], Rodrigo Mahu [a], Fernando Huenupan [c], Néstor Becerra Yoma [a,*]

[a] *Speech Processing and Transmission Lab., Electrical Engineering Department, University of Chile, Av, Tupper 2007, Chile*
[b] *Multimodal Signal Processing (MSP) Lab, Department of Electrical and Computer Engineering, The University of Texas at Dallas, Richardson TX 75080, USA*
[c] *Department of Electrical Engineering, Universidad de La Frontera, Francisco Salazar 01145, Temuco, Chile*

ARTICLE INFO

Keywords:
Speech emotion recognition
Human-robot interaction
Static and dynamic conditions
Acoustic modelling
Beamforming responses

ABSTRACT

The use of speech-based solutions is an appealing alternative to communicate in human-robot interaction (HRI). An important challenge in this area is processing distant speech which is often noisy, and affected by reverberation and time-varying acoustic channels. It is important to investigate effective speech solutions, especially in dynamic environments where the robots and the users move, changing the distance and orientation between a speaker and the microphone. This paper addresses this problem in the context of speech emotion recognition (SER), which is an important task to understand the intention of the message and the underlying mental state of the user. We propose a novel setup with a PR2 robot that moves as target speech and ambient noise are simultaneously recorded. Our study not only analyzes the detrimental effect of distance speech in this dynamic robot-user setting for speech emotion recognition but also provides solutions to attenuate its effect. We evaluate the use of two beamforming schemes to spatially filter the speech signal using either delay-and-sum (D&S) or minimum variance distortionless response (MVDR). We consider the original training speech recorded in controlled situations, and simulated conditions where the training utterances are processed to simulate the target acoustic environment. We consider the case where the robot is moving (dynamic case) and not moving (static case). For speech emotion recognition, we explore two state-of-the-art classifiers using hand-crafted features implemented with the ladder network strategy and learned features implemented with the wav2vec 2.0 feature representation. MVDR led to a signal-to-noise ratio higher than the basic D&S method. However, both approaches provided very similar average concordance correlation coefficient (CCC) improvements equal to 116 % with the HRI subsets using the ladder network trained with the original MSP-Podcast training utterances. For the wav2vec 2.0-based model, only D&S led to improvements. Surprisingly, the static and dynamic HRI testing subsets resulted in a similar average concordance correlation coefficient. Finally, simulating the acoustic environment in the training dataset provided the highest average concordance correlation coefficient scores with the HRI subsets that are just 29 % and 22 % lower than those obtained with the original training/testing utterances, with ladder network and wav2vec 2.0, respectively.

[*] Corresponding author.
E-mail address: nbecerra@ing.uchile.cl (N.B. Yoma).

https://doi.org/10.1016/j.csl.2024.101666
Received 19 September 2023; Received in revised form 10 May 2024; Accepted 21 May 2024
Available online 22 May 2024

1. Introduction

Seamless human-robot collaboration will be a strategic component in commercial applications in the next 10 to 20 years. Consequently, social robotics is one of the most important and critical challenges in robotic science and engineering (Stock-Homburg, 2022). Although some progress has been made on this topic, comprehensive social interaction between humans and robots in real-world conditions is not possible today. 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." (Yang et al., 2018). To achieve this purpose, 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, and can raise privacy concerns. 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. In this context, one of the most challenging and least explored scenarios is that in which one or more users, who may be in motion, attempt to interact with the robot, which may also be in motion. In addition, this interaction may occur in a noisy environment, which affects this communication.

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 to adapt their behavior (Rossi et al., 2017). In addition, social profiling also refers to the ability to recognize social phenomena, such as commitment, conflict, empathy, interest, and emotions, which cannot be directly observed but must be inferred by examining indirect indicators. Some of these indirect indicators can be body posture (Gaschler et al., 2012), facial expressions (Faria et al., 2017; Deng et al., 2019), gaze direction (Paletta et al., 2019; Chakraborty et al., 2021), voice volume, etc. We are particularly interested in Speech Emotion Recognition (SER), which seeks to dynamically detect the emotional state of the user during the interaction using acoustic features (Busso et al., 2013; Scherer, 2003). Within social user profiling, the concept of emotion recognition arises because the user may exhibit multiple emotions during the interaction with the robot.

The vast majority of the research in this discipline is focused on human-computer interaction (HCI) (Shah Fahad et al., 2021), assuming the user is directly next to the microphone. In this scenario, the influence of the acoustic channel is neglected since the speech is clean. Only a few studies have tested distant SER in real environments (Shah Fahad et al., 2021; Chen et al., 2020; Leem et al., 2021). 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 Salekin et al. (2017), it is proposed to select 48 robust 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. Spectral and temporal filtering was performed. However, no beamforming technique was used, so it is expected that better results may be achieved by combining the audio from the microphones. In Ahmed et al. (2017), the use of a metric to determine the distortion of the features according to the distance to the microphone was suggested. In addition, they trained their classifier with convoluted audio with artificially generated room impulse responses (RIRs). The proposed solution used the weighted prediction error (WPE) algorithm to remove reverberation from the test audios and Coherent-to-Diffuse Power Ratio Estimation (CDR) to perform de-noise. However, in this study only static situations are evaluated, varying the distance to the microphone. In Chen et al. (2020), a feature acquisition technique using a robotic platform with a Kinect mounted on the robot was evaluated. The test database is acted by the volunteers from their research lab and has only 500 utterances. Furthermore, the authors neither use any speech enhancement technique nor evaluate the robot motion effect. However, a dynamic moving scenario between the robot and the user is important to consider, since robots are crucial both in industrial tasks (Berg et al., 2019; Kousi et al., 2019) and in butler or personal assistant tasks (Chen et al., 2020; Miseikis et al., 2020). Although there is consensus on the importance of mobile HRI, there are few studies that have analyzed the effect of this scenario on the voice channel for speech-based systems relying on the voice as their main input.

This paper explores the recognition of emotional speech in HRI under a challenging, but realistic scenario where the distance and angle between the robot and the user vary through the interaction. Our study not only analyzes the detrimental effect of distance speech in this dynamic robot-user setting for SER tasks but also provides solutions to attenuate its effect. While there are studies on the effect of dynamism in HRI for the speech-to-text task (Novoa et al., 2021), the performance of SER models in mobile HRI scenarios has not been tested so far. Our evaluation relies on the MSP-Podcast corpus (Lotfian and Busso, 2019), which is currently the largest naturalistic emotional dataset in the community. This database, unlike most other databases (Busso et al., 2008; Metallinou et al., 2016; Cao et al., 2014), contains fragments of non-acted audio, in normal speech environments, so it better matches the expressive behaviors observed in real-world interactions. We propose a setup for re-recording the test partition of our database. The setup includes a meeting room with a PR2 robot and three loudspeakers. One of the loudspeakers plays the target speech signal, while the other two loudspeakers reproduce ambient noise. The loudspeakers are positioned two meters from the robot's central position of movement. Specifically, the target speech loudspeaker is placed directly in front of the robot, while the noise speakers are positioned 45° apart from the speech speaker, with one on each side. Additionally, the robot has the capability to move one meter forward and one meter away from the target source from the central position. Simultaneously, the robot head also rotates. 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.

Our solution to mitigate the channel modeling problem includes the use of beamforming and RIR. We simulate the 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

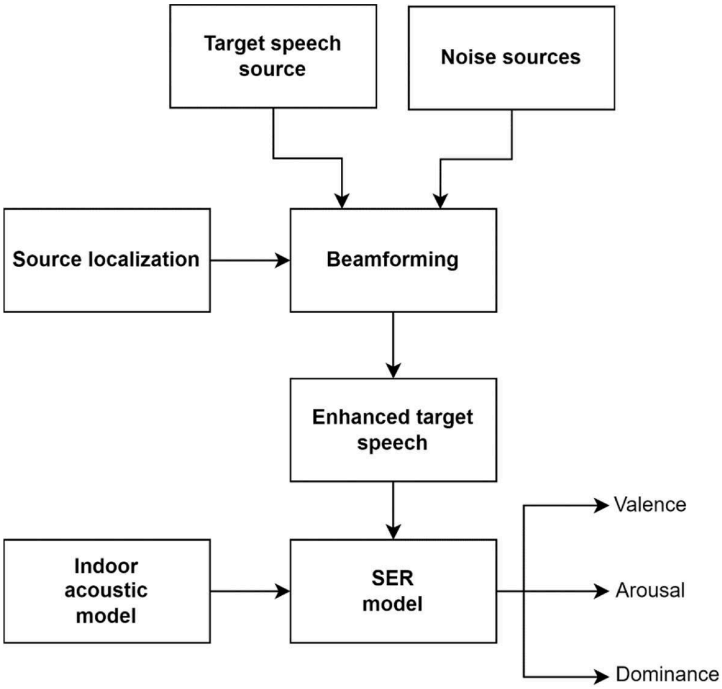

Fig. 1. Proposed SER system for static and dynamic indoor HRI scenarios. Source localization is considered information provided by, for instance, image processing for both target and noise sources to be used by a given beamforming technology. The enhanced speech is input to the SER classifier, which in turn is trained with the target indoor acoustic model. Finally, the SER classifier delivers the emotional attributes.

array lobe. 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 (Simmer et al., 2001). Some studies (Novoa et al., 2021; Díaz et al., 2021) have compared different beamforming techniques for an automatic speech recognition (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: the delay-and-sum (D&S) scheme (Omologo et al., 2001) and the minimum variance distortionless response (MVDR) (Bitzer and Simmer, 2001). We also address the acoustic channel modeling problem by using RIRs to simulate a real environment in the training database.

At the model architecture level, we reduce the mismatch between the train and test conditions by using a semi-supervised domain adaptation strategy. One of the most popular architectures in SER is the ladder network, especially those using semi-supervised training (Huang et al., 2018; Parthasarathy and Busso, 2018; Parthasarathy and Busso, 2020; Tao et al., 2019). This type of network consists of an encoder-decoder scheme with lateral connections between these two modules. The encoder is trained to perform classification or regression tasks depending if the reference corresponds to emotion classes or attributes, respectively, with an input to which noise, usually Gaussian, is added at each of the layers. The decoder is trained to perform a reconstruction of the original input (before adding the noise) of each layer. The feature extraction procedure is composed of two stages. First, frame-by-frame low-level descriptors (LLDs) are retrieved. Mel frequency cepstral coefficients (MFCCs), fundamental frequency (F0), and energy are all included in the set. Over the LLDs, several statistics known as high-level descriptors (HLDs) are computed creating a 6373-dimensional feature vector, regardless of the length of the sentence. This set is then passed to the ladder network. While this strategy has been adapted for SER in noisy speech (Leem et al., 2021), this study evaluates the approach in a more complex mobile HRI scenario. We formulate the SER problem as a regression task in this paper, where we estimate the emotional attributes of arousal, dominance, and valence. Emotional attributes reduce the complexity of describing human emotions using categorical labels with a few meaningful dimensions (Devillers et al., 2005; Mower et al., 2009).

This method is a first step towards a more complete integration of SER in HRI, including static and dynamic situations. The main contribution of this study is the proposed experimental setting to simulate real dynamic HRI scenarios, and the evaluation of state-of-the-art SER solutions and techniques for noise robustness to address this challenging problem that has not been addressed elsewhere. An important contribution is also the testing dataset re-collected using the proposed robotic platform, which in turn can be shared with the community. According to the strategy followed here, recording training speech in several operating HRI conditions would not be necessary. Instead, we model the indoor testing environment from the reverberation, noise, and also beamforming response points of view to simulate the testing condition in the training data. The procedure employed and studied here in the framework of SER is interesting because it allows us to cope with the problem of complex real HRI scenarios with limited training data. We hypothesize that the critical factors influencing SER in the HRI scenarios with limited training data that are targeted here correspond to: variability in acoustic conditions represented with additive noise and reverberation; the increase of interference caused by additive noise and reverberation when the speaker-microphone distance increases from a few centimeters to more than one meter; the response of the beamforming schemes; and the time-varying acoustic channel effect that is observed in dynamic conditions (Novoa et al., 2021). The approach adopted here targets these problems.

2. Proposed framework

Our paper builds upon our preliminary study (Grageda et al., 2023). 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 (Díaz et al., 2021). By doing so, it is possible to avoid the error introduced by reverberation in indoor scenarios. In contrast to Grageda et al. (2023) where results with only simulated and real static conditions were presented, in this paper a much more complex HRI scenario was incorporated to include translational and rotational robot movement. This dynamic condition has not been addressed elsewhere in the field of SER. Also, in addition to Grageda et al. (2023), the static and dynamic HRI scenarios considered here were evaluated with the transfer learning methodology by using an architecture based on wav2vec 2.0 (Wagner et al., 2023). We used the foundation model (without emotion-specific fine-tuning) and carried out the fine-tuning by ourselves following the same procedure described in Wagner et al. (2023).

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) (Novoa et al., 2021). The following assumptions are included in this framework: first, the angular position of the target source can be accurately estimated 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 presented in Novoa et al. (2021).

Two beamforming techniques are considered in this study: D&S and MVDR (see Appendix). 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 this paper, indoor acoustic modeling (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 modeled similarly to Novoa et al. (2021), 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 real HRI conditions more accurately.

Two neural network-based schemes were evaluated in this study. The first architecture employed here corresponded to the ladder network proposed in Parthasarathy and Busso (2020). The network is trained with multitask learning, jointly predicting arousal, valence, and dominance. The input to the network is the ComParE feature set (Schuller et al., 2013), which has 6373 HLDs, regardless of the audio duration of the speech segment. For training, we ran 100 epochs with a learning rate set to 0.0001. We used the same code provided by the authors in Parthasarathy and Busso (2020) in our implementation to obtain our results.

The second architecture tested in this study was the approach with the highest concordance correlation coefficient (CCC) in the SER tasks using the MSP-Podcast corpus reported in Wagner et al. (2023). This model has a simple head on top of a pre-trained wav2vec 2.0-large-robust transformer (Hsu et al., 2021). An average pooling was applied to the hidden states of the last transformer layer, passing them through a fully connected layer of 1024 nodes and a final output layer of 3 nodes. A dropout procedure was employed before the two head layers. Before fine-tuning the model, the weights of the Convolutional Neural Network layers were frozen, but the transform and head layers were trained as suggested in Wang et al. (2021). Adaptive Moment Estimation (ADAM) optimizer was applied with the learning rate and batch size equal to 10^{-4} and 32, respectively. The wav2vec 2.0 transformer estimates 1024 features per window. Then, the pooling layer computes the global average per each feature, resulting in a vector of dimension 1024 that is input to the hidden layers of the head. We used the codes shared by the authors of Wagner et al. (2023) to obtain our results.

2.2. Robotic platform and recording settings

We employed the publicly available MSP-Podcast corpus (Lotfian and Busso, 2019) (version 1.9), collected by the Multimodal Signal Processing Laboratory at the University of Texas in Dallas. The corpus is collected from available recordings from audio-sharing websites, including natural conversations about a broad range of topics. Therefore, the recordings are good representations of speech expected to be collected in real-world applications. We chose to train and evaluate our models with the MSP-Podcast corpus since it is currently one of the largest naturalistic databases available in the community. The recordings are obtained from real human interactions. Given that our focus is on exploring solutions for real human-robot interactions, it is natural to use this database instead of other naturalistic or acted corpora. Common benchmark databases with acted recordings such as IEMOCAP, EMO-DB, MSP-Improv, and CREMA-D are recorded from a few speakers. Instead, the MSP-Podcast includes recordings from 1354 speakers. The speaker diversity also makes this corpus a natural option for our experiments. It has 86,389 speech turns, corresponding to 137 h 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

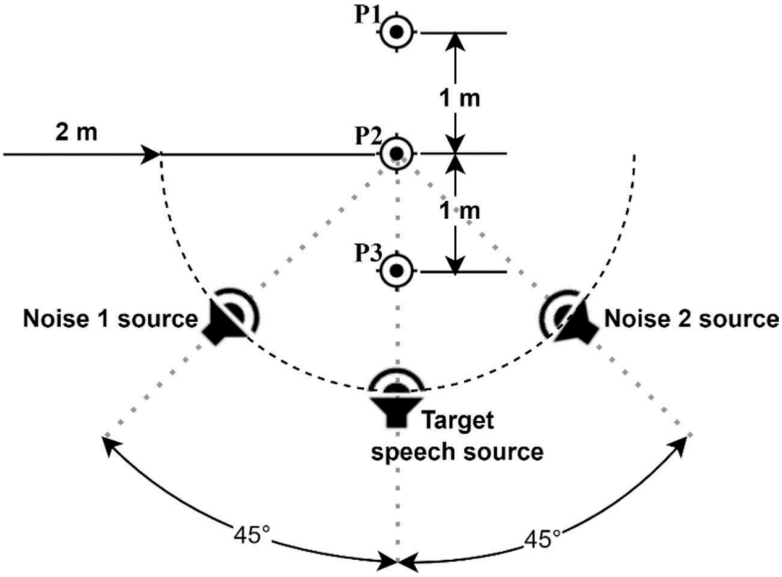

Fig. 2. Schematic view of the testbed.

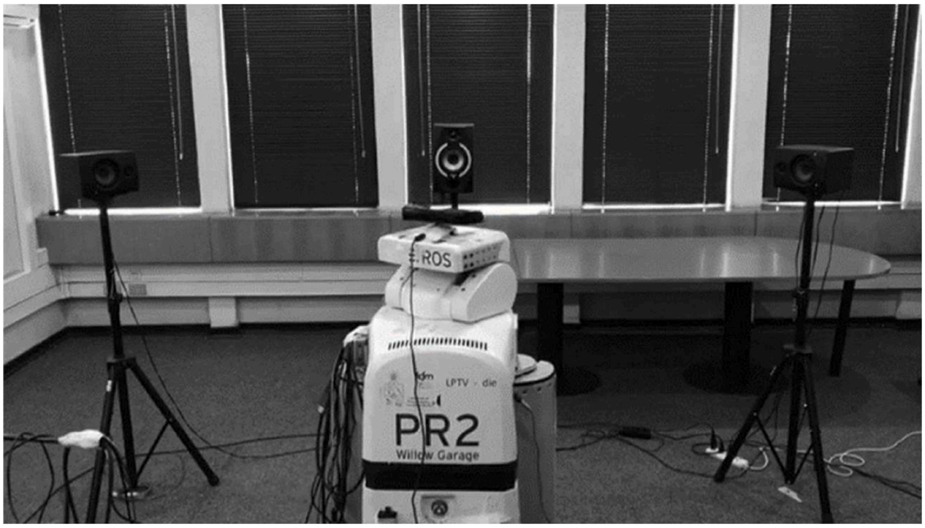

Fig. 3. Rear view of the testbed.

annotated via a modified version of the crowdsourcing protocol presented in Burmania et al. (2016). This study focuses on the estimation of arousal (calm to active), valence (negative to positive), and dominance (weak to strong), formulating the task as a regression problem.

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 h of audio. The HRI testbed was implemented with our Personal Robot 2 (PR2) robot equipped with a Microsoft Xbox 360 Kinect sensor mounted on top of its head. As shown in Fig. 2, we use one target speech and two noise sources, each one located two meters 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 5 dB measured at point *P2*. Figs. 3 and 4 show the meeting room with the robot and the studio speakers.

We have two scenarios: static and dynamic settings. For the static scenario, the PR2 robot stays still at *P2*, with its head pointing directly to the speech source. In the dynamic scenario, PR2 moves between P1 and P3 at a speed of 0.45 m/s. Moreover, the head of the robot moves periodically between 50° and -50° with a uniform angular velocity of 0.56 rad/s, simulating following a moving target as shown in Fig. 5. While recording the dynamic scenario, the angles that the PR2′s head are stored in real-time. These angles are then used to calculate the DOA corresponding to the target speech.

In contrast to the approach proposed in Novoa et al. (2021), three sets of 63 RIR per each Microsoft Kinect microphone were

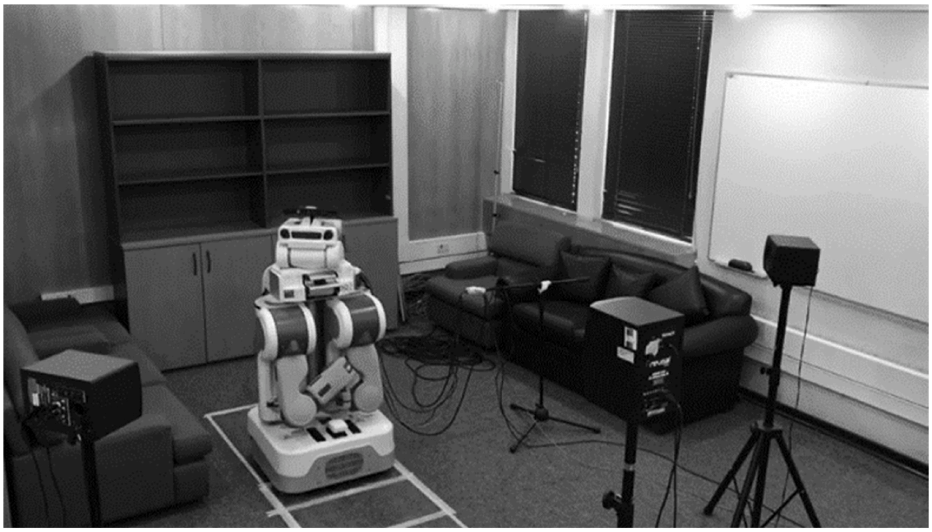

Fig. 4. Side view of the testbed.

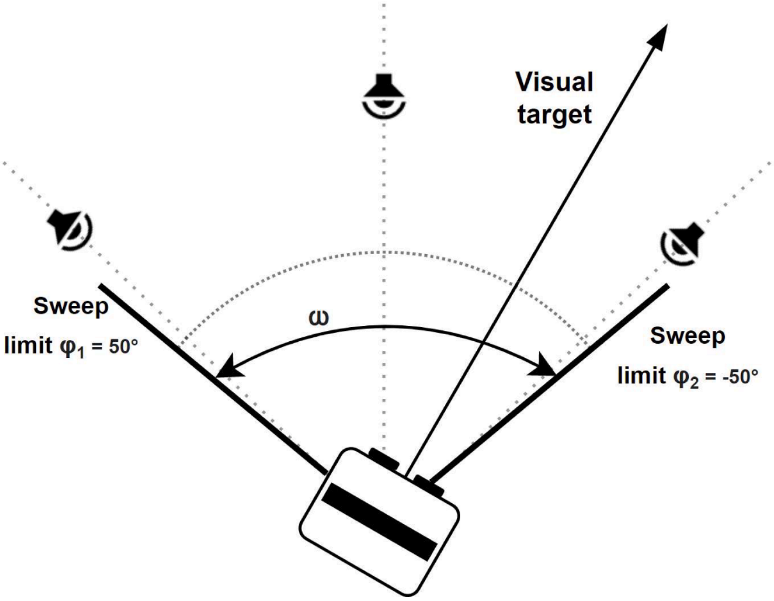

Fig. 5. Robot head movement.

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 (2000). An exponential sweep from 64 Hz to 8 kHz sine functions was generated and played back with a studio loudspeaker located at the 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. Training and testing databases

3.1. Training datasets

The SER architectures evaluated here were trained with two types of data. First, we used the original MSP-Podcast corpus, which

Table 1
Valence, arousal, and dominance when the ladder network and wav2vec-based classifiers were trained with the original MSP-Podcast training dataset and tested with the static HRI subset. The result with the original MSP-Podcast testing dataset is included as a reference.

Model	Test type	SNR [dB]	CCC Aro	CCC Dom	CCC Val	Mean CCC
Ladder network	*Original Testing Data*	–	0.571	0.461	0.216	0.416
	HRI Static Data	5.46	0.172	0.112	0.042	0.109
	HRI Static Data + D&S	8.34	0.328	0.295	0.062	0.229
	HRI Static Data + MVDR	9.35	0.358	0.305	0.078	0.247
Wav2vec	*Original Testing Data*	–	0.606	0.500	0.518	0.541
	HRI Static Data	5.46	0.435	0.367	0.242	0.348
	HRI Static Data + D&S	8.34	0.471	0.409	0.313	0.398
	HRI Static Data + MVDR	9.35	0.424	0.336	0.268	0.343

we refer to as *Original Training Dataset*. 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 refer to this setting as *Simulated Training Dataset*.

A simulated training dataset was generated with 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 looked directly at 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, noise was artificially added to the resulting audios at SNRs that were randomly chosen between 10 dB and 20 dB. The additive noise was obtained as follows: noise segments from the DEMAND database (Thiemann et al., 2013) were convoluted with the impulse responses from *RIR-Noise1 Source* and *RIR-Noise2 Source*; then, they were added with a ratio equal to one; and, the resulting external additive noise was summed to the PR2 engine noise at SNRs between -5 dB and 5dB. Observe that the speech and noise RIRs employed to generate a given noisy signal correspond to the same robot and head positions. 21 angular head positions were employed, where each head position determines a given DOA. Moreover, the resulting reverberated noisy data from the four Microsoft Kinect microphones were delayed and combined with the D&S and MVDR beamforming methods producing two training sets that simulate the target acoustic environment: *Simulated Training Data+D&S* and *Simulated Training Data+MVDR*, respectively.

3.2. Testing databases

Results with four testing conditions are reported here: *Original Testing Data*, corresponding to the clean audios from the test partition of the MSP-Podcast corpus; testing data that simulates the target acoustic environment corresponding to the audios from the test partition of the MSP-Podcast corpus which were processed similarly to the training data (Section 3.1), *Simulated Testing Data; HRI Static Data*, corresponding to testing audios from the MSP-Podcast corpus re-recorded in the robotic platform in static conditions (see Section 2.2); and *HRI Dynamic Data*, corresponding to testing audios from the MSP-Podcast corpus re-recorded in the robotic platform with dynamic conditions (see Section 2.2). Beamforming schemes D&S and MVDR were assessed with the following conditions: *Simulated Testing Data* (i.e. *Simulated Testing Data+D&S* and *Simulated Testing Data+MVDR*), *HRI Static Data (*i.e. *HRI Static Data+D&S* and *HRI Static Data+MVDR),* and *HRI Dynamic Data* (i.e. *HRI Dynamic Data+D&S* and *HRI Dynamic Data+MVDR*).

4. Results and discussion

The ladder network and wav2vec 2.0 strategies were trained ten times and twice, respectively, in each experiment. It should be noted that to train the wav2vec 2.0 architecture requires much more time than the ladder network one. About the significance analysis, the MSP-Podcast corpus is large enough and the original test subset was randomly divided into 15 smaller subsets, where each subset contains more than 1.000 utterances (a thousand samples is much more than the full size of most emotional datasets available in the literature). This size is maintained to preserve the statistical robustness and reliability of our analysis. This approach allows us to perform a detailed statistical analysis to validate our findings. Specifically, we randomly split the original test set into 15 similarly sized subsets and reported the average results from these subsets. This method was implemented to facilitate a two-tailed *t*-test across the 15 test subsets, aiming to ensure that our findings are statistically robust and significant. We defined statistical significance at a p-value of 0.05, following standard practices in the field. It is important to note that this subdivision and analysis strategy does not involve the replacement of samples, which distinguishes it from a bootstrapping methodology. The approach that we adopted ensures that each subset is unique and that there is no overlap or repetition of utterances across subsets. This method was chosen to maintain the integrity of the test set and to ensure that our statistical analysis reflects genuine performance variations across different parts of the dataset. This procedure has been previously employed and validated in Lin and Busso (2023), where it was found that it offers a robust framework for evaluating model performance across diverse data segments.

4.1. Original training data & real HRI testing scenarios

Table 1 and Fig. 6 present the results when the ladder network and wav2vec 2.0 were trained with the *Original Training Data* set and tested with static testing scenarios. The testing subsets corresponded to *Original Testing Data, HRI Static Data, HRI Static Data+D&S,* and *HRI Static Data+MVDR*. According to Table 1, the highest CCC degradation in arousal, dominance, and valence when compared with

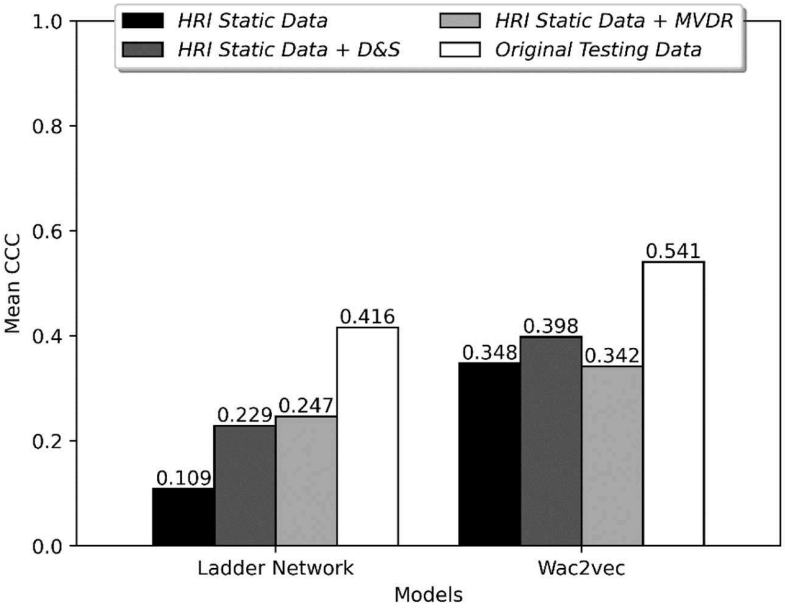

Fig. 6. Average CCC across valence, arousal and dominance according to Table 1. The ladder network and wav2vec based classifiers were trained with the original MSP-Podcast training dataset and tested with the static HRI subset. The result with the original MSP-Podcast testing dataset is included as a reference.

Table 2
Valence, arousal, and dominance when the ladder network and wav2vec-based classifiers were trained with the original MSP-Podcast training dataset and tested with the dynamic HRI subset. The result with the original MSP-Podcast testing dataset is included as a reference.

Model	Test type	SNR [dB]	CCC Aro	CCC Dom	CCC Val	Mean CCC
Ladder Network	*Original Testing Data*	–	0.571	0.461	0.216	0.416
	HRI Dynamic Data	5.69	0.176	0.093	0.052	0.107
	HRI Dynamic Data + D&S	8.37	0.339	0.288	0.066	0.231
	HRI Dynamic Data + MVDR	10.82	0.332	0.267	0.077	0.225
Wav2vec	*Original Testing Data*	–	0.606	0.500	0.518	0.541
	HRI Dynamic Data	5.69	0.426	0.356	0.268	0.350
	HRI Dynamic Data + D&S	8.37	0.449	0.385	0.297	0.377
	HRI Dynamic Data + MVDR	10.82	0.434	0.339	0.304	0.359

the *Original Testing Data* results corresponds to the *HRI Static Data set* with the ladder network. The degradation with wav2vec 2.0 was much smaller. This result must be because this model was pre-trained using 2k hours of noisy data (Hsu et al., 2021). The beamforming schemes D&S and MVDR increased the SNR and decreased the degradation in CCC for arousal, dominance, and valence when compared with the *Original Testing Data* results using the ladder network. The increase in SNR was equal to 53 % and 71 % with the D&S and MVDR strategies, respectively. As can be seen in Fig. 6, when compared with set *HRI Static Data*, D&S, and MVDR led to an increase in the average CCC equal to 110 % and 127 %, respectively, when using the ladder network. Surprisingly, only D&S could increase the average CCC when wav2vec 2.0 was employed (14 %). Moreover, this improvement was considerably smaller than with the ladder network. This result should also be because, as aforementioned, wav2vec 2.0 was also exposed to noisy data during the pre-training procedure, so the room for improvement due to spatial filtering is smaller. Observe that, even though MVDR provides a higher SNR improvement than D&S, it does not necessarily lead to higher improvements in average CCC than the latter. This could be due to the artifacts introduced by MVDR (Erdogan et al., 2016).

Table 2 and Fig. 7 show the results when ladder network and wav2vec 2.0-based models were trained with the *Original Training Data set* and tested with dynamic testing scenarios. The testing subsets corresponded to *Original Testing Data* (as a reference), *HRI Dynamic Data, HRI Static Data+D&S,* and *HRI Dynamic Data+MVDR*. According to Table 2, the highest CCC degradation in arousal, dominance, and valence when compared with the results with *Original Testing Data* is also observed with the *HRI Dynamic Data* set using the ladder network approach. As expected, the wav2vec 2.0-based engine is more robust than the ladder network one because it is pre-trained with noisy data. Consequently, beamforming methods D&S and MVDR led to smaller improvements in CCC for all the attributes with wav2vec 2.0 than with the ladder network architecture. The beamforming schemes D&S and MVDR increased the SNR and reduced the CCC degradation for all the emotional attributes when compared with the *Original Testing Data* when using the ladder network classifier. Observe that the improvements in SNR are similar to those in Table 1. According to Fig. 7 and similarly to Fig. 6, D&S and MVDR led to an increase in the average of CCCs equal to 116 % and 110 %, respectively, when using the ladder network.

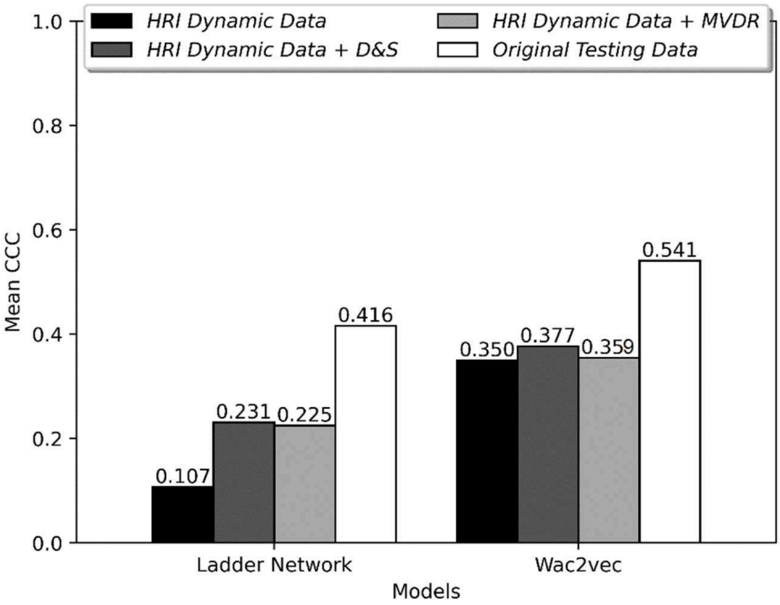

Fig. 7. Average CCC across valence, arousal and dominance according to Table 2. The ladder network and wav2vec-based classifiers were trained with the original MSP-Podcast training dataset and tested with the dynamic HRI subset. The result with the original MSP-Podcast testing dataset is included as a reference.

However, Fig 7 shows that D&S could lead to a greater increase in the average CCC than MVDR when the wav2vec 2.0-based model was employed although this relative improvement was smaller than with the ladder network classifier. These results corroborate the hypothesis mentioned above about the potential artifacts introduced by the MVDR method even though it can lead to higher reductions in SNR than D&S.

In contrast to what was observed with ASR in a similar HRI scenario (Novoa et al., 2021, 2018), the SER technology evaluated here showed a baseline degradation that is quite similar for both the static (Table 1 and Fig. 6) and dynamic (Table 2 and Fig. 7) scenarios. In ASR, the search is carried out on a frame-by-frame basis. Consequently, if the severity of the acoustic conditions is time-dependent, the ASR search in those frames with lower local SNR or higher reverberation effect may condition the search in the following frames. However, as discussed above, the ladder network and wav2vec 2.0-based schemes deliver the emotional attributes by extracting features on an utterance-by-utterance basis. Moreover, the periods of the robot's translational and rotational movements are approximately equal to nine seconds and seven seconds, respectively, which in turn are comparable to the average length of the utterances (i.e. 5.5 s). As a result, there could be an acoustic severity compensation effect along the whole utterances to estimate the emotional attributes.

Valence is more dependent on linguistic information than arousal and dominance (Wagner et al., 2023). The ASR accuracy degradation in a similar HRI scenario (Novoa et al., 2021, 2018) explains the greater CCC reduction for valence than for arousal and dominance observed here. It is worth mentioning that the ladder network delivers the emotional attributes on an utterance-by-utterance basis employing statistics computed over short-term handcrafted features. In contrast, wav2vec 2.0 extracts the features from the raw signal using CNN layers. The task employed to train wav2vec 2.0 makes this representation implicitly capture linguistic information. The authors in Wagner et al. (2023) demonstrated this observation by synthesizing neutral speech with transcription of the original emotional sentences. Even though the acoustic properties were neutral, the model was able to classify valence with acceptable performance. In other words, wav2vec 2.0 depends more on the linguistic information extracted from the input than the ladder network. As a consequence, one can expect that the relative degradation of the performance for valence should be greater than the ones for arousal and dominance. As can be seen in Tables 1 and 2, when the ladder network was trained with the original MSP-Podcast training dataset, the average relative degradations in CCC for arousal, dominance, and valence were 69.5 %, 77.7 %, and 78.2 %, respectively, when the original MSP-Podcast testing dataset was replaced with the static and dynamic HRI data. Under the same conditions, wav2vec 2.0 provided CCC reductions for arousal, dominance, and valence equal to 29.0 %, 27.7 %, and 50.8 %, respectively.

4.2. *Models trained and tested with simulated data*

Table 3 and Fig. 8 show the results when the ladder network and wav2vec 2.0-based classifiers were trained and tested with simulated data (described in Section 2.1). Two training/testing conditions were employed: *Simulated Data+D&S* and *Simulated Data+MVDR*, where D&S and MVDR were applied to the data that simulates the acoustic environment, as explained in Section 3.1. Results with the *Simulated Data+D&S* and *Simulated Data+MVDR* sets correspond to the static conditions, i.e. the robot movement is

Table 3
Valence, arousal, and dominance when the ladder network and wav2vec-based classifiers were trained and tested with the original MSP-Podcast dataset that had incorporated the acoustic model of the test room plus the response of the beamforming schemes evaluated here, i.e. D&S and MVDR.

Model	Training database	Testing database	CCC Aro	CCC Dom	CCC Val	Mean CCC
Ladder Network	*Simulated Training Data + D&S*	*Simulated Testing Data + D&S*	0.496	0.363	0.113	0.324
	Simulated Training Data + MVDR	*Simulated Testing Data + MVDR*	0.482	0.350	0.101	0.311
Wav2vec	*Simulated Training Data + D&S*	*Simulated Testing Data + D&S*	0.556	0.442	0.303	0.434
	Simulated Training Data + MVDR	*Simulated Testing Data + MVDR*	0.513	0.402	0.203	0.373

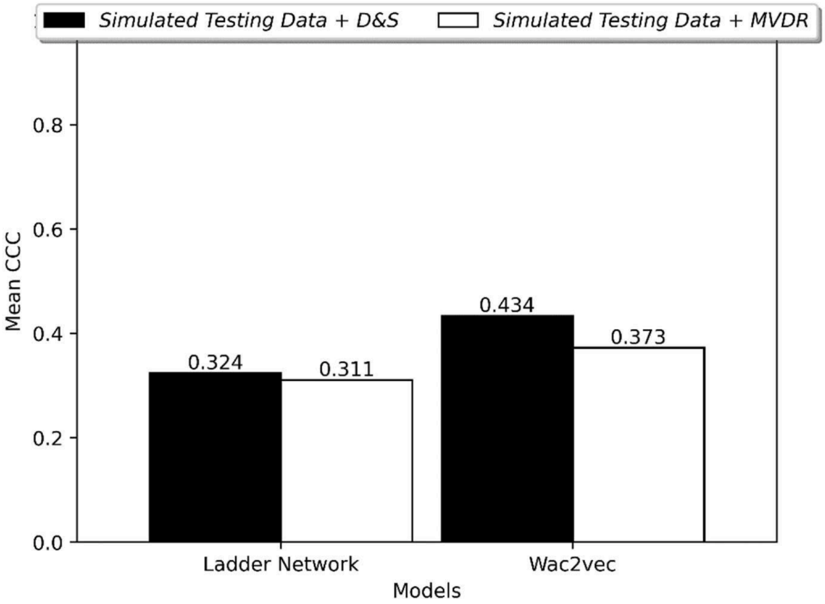

Fig. 8. Average CCC across valence, arousal, and dominance according to Table 3. The ladder network and wav2vec-based classifiers were trained and tested with the MSP-Podcast database that had the acoustic model incorporated in the utterances according to Section 3.1.

Table 4
Valence, arousal, and dominance when the ladder network and wav2vec-based classifiers were trained with the original MSP-Podcast training partition that had incorporated the acoustic model of the test room plus the response of the beamforming schemes evaluated here, i.e. D&S and MVDR. The testing subset corresponded to the testing partition of MSP-Podcast that was re-recorded in static condition with the HRI testbed according to Section 2.2.

Model	Train database	Test	CCC Aro	CCC Dom	CCC Val	Mean CCC
Ladder Network	*Simulated Training Data + D&S*	*HRI Static Data + D&S*	0.407	0.317	0.093	0.272
	Simulated Training Data + MVDR	*HRI Static Data + MVDR*	0.440	0.341	0.099	0.293
Wav2vec	*Simulated Training Data + D&S*	*HRI Static Data + D&S*	0.556	0.458	0.320	0.445
	Simulated Training Data + MVDR	*HRI Static Data + MVDR*	0.546	0.406	0.258	0.403

not emulated, and can be compared with the results obtained with the *Original Training Data/Original Testing Data sets* in Tables 1 or 2. According to Table 3, *Simulated Data+D&S* and *Simulated Data+MVDR* with ladder network still led to reductions in the average CCC score equal to 22 % and 25 %, respectively, when compared with the results with the *Original Training Data/Original Testing Data sets* in Tables 1 or 2. Although the training and testing conditions in Table 3 are intended to be matched, the added noise and reverberation still introduce some uncertainty. Even in this case, the average CCC value achieved by the ladder network architecture is 42 % and 26 % greater than those with testing subsets *HRI Static Ddata+D&S* and *HRI Static Data+MVDR* in Table 1 where the training/testing mismatch is highly significant.

As can be seen in Tables 1 and 3, *Simulated Data+D&S* with wav2vec 2.0 led to a reduction in the average of the CCC scores as small as 20 % when compared with the *Original Training Data/Original Testing Data* results. Moreover, the average CCC score is 9 % higher than the one with the *HRI Static Data+D&S* set. These results suggest that the wav2vec 2.0-based classifier is more robust to the training/testing mismatch condition than the ladder network-based model. *Simulated Data+MVDR* with wav2vec 2.0 experiments led to an improvement in the average CCC score equal to 9 % when compared with the *HRI Static Data+MVDR* results.

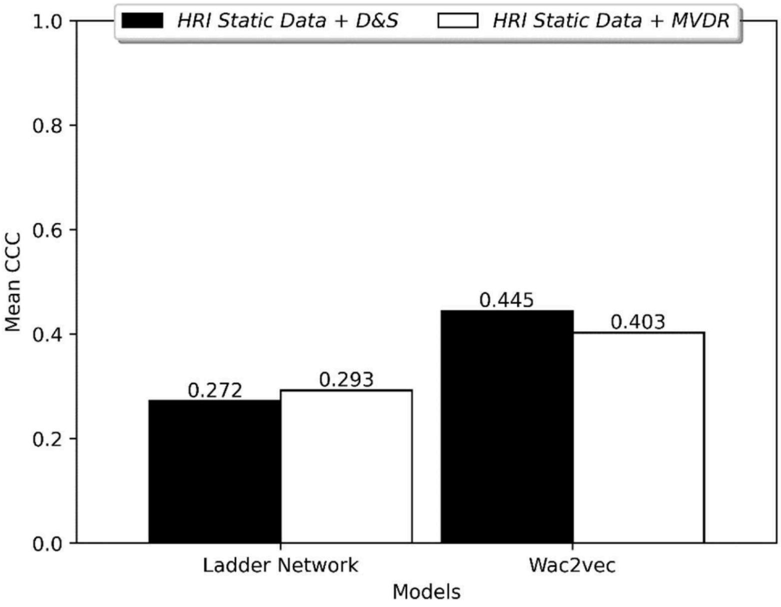

Fig. 9. Average CCC across valence, arousal and dominance according to Table 4. The ladder network and wav2vec-based classifiers were trained with the MSP-Podcast training subset that had the acoustic model incorporated in the utterances according to Section 3.1. The testing subset corresponded to the testing partition of MSP-Podcast that was re-recorded in static condition with the HRI testbed according to Section 2.2.

Table 5
Valence, arousal, and dominance when the ladder network and wav2vec-based classifiers were trained with the original MSP-Podcast training partition that had incorporated the acoustic model of the test room plus the responses of the beamforming schemes evaluated here, i.e. D&S and MVDR. The testing subset corresponded to the testing partition of MSP-Podcast that was re-recorded in the dynamic condition with the HRI testbed according to Section 2.2.

Model	Train type	Test type	CCC Aro	CCC Dom	CCC Val	Mean CCC
Ladder Network	*Simulated Training Data + D&S*	*HRI Dynamic Data + D&S*	0.473	0.350	0.109	0.311
	Simulated Training Data + MVDR	*HRI Dynamic Data + MVDR*	0.461	0.344	0.098	0.301
Wav2vec	*Simulated Training Data + D&S*	*HRI Dynamic Data + D&S*	0.551	0.455	0.299	0.435
	Simulated Training Data + MVDR	*HRI Dynamic Data + MVDR*	0.552	0.416	0.277	0.415

4.3. Models trained with simulated data and tested in real HRI scenarios

Table 4 and Fig. 9 present the results when the ladder network and wav2vec 2.0-based architectures were trained with data that simulates the acoustic environment and tested with real static HRI data. As can be seen in Tables 3 and 4, the difference between the average CCC scores obtained with the *HRI Static Data+D&S* and *Simulated Testing Data+D&S* sets when the ladder network was trained with *Simulated Training Data+D&S* was 16 %. A similar result was observed with the *HRI Static Data+MVDR* and *Simulated Testing Data+MVDR* sets when the difference in the average CCC metrics was only 6 %. The same trend was observed with the wav2vec 2.0-based model, although the average CCC scores were slightly greater with *HRI Static Data+MVDR* set than with the *Simulated Testing Data+MVDR* set. This difference might be due to a random behavior of the artifact introduced by MVDR or by the variability in SNR provided in the *Simulated Testing Data+MVDR set.*

Results in Table 4 and Fig. 9 basically suggest that the training conditions proposed here, that simulate the target acoustic environment, are reasonable approximations to the real static HRI scenario. This result is corroborated in Table 5 and Fig. 10 which present the results when the ladder network and wav2vec 2.0-based architectures were trained with *Simulated Training Data+D&S* and *Simulated Training Data+MVDR* and tested with the real dynamic HRI subset. The CCC metrics obtained with the dynamic HRI environments and the model trained with simulated data (Table 5 and Fig. 10) are similar to those achieved in static conditions with the same trained models (Table 4 and Fig. 9). First, this result validates the acoustic modeling-based scheme to generate the training data that emulates the target scenario. Particularly, it is worth emphasizing the incorporation of a suitable variability of DOAs and loudspeaker-robot distance. Second, as observed earlier, this result suggests that the SER schemes evaluated here were naturally robust to the dynamic conditions employed in our robotic platform. Actually, the average CCC score was slightly higher in the dynamic HRI condition than in the static one. This is probably because the beamforming effectiveness is not homogenous and depends on both DOA and the audio speaker-robot distance. Also, the wav2vec 2.0-based model provided average CCCs that are 50 % and 39 % higher than those achieved with the ladder network-based model for static and dynamic scenarios, respectively. This may be because the wav2vec

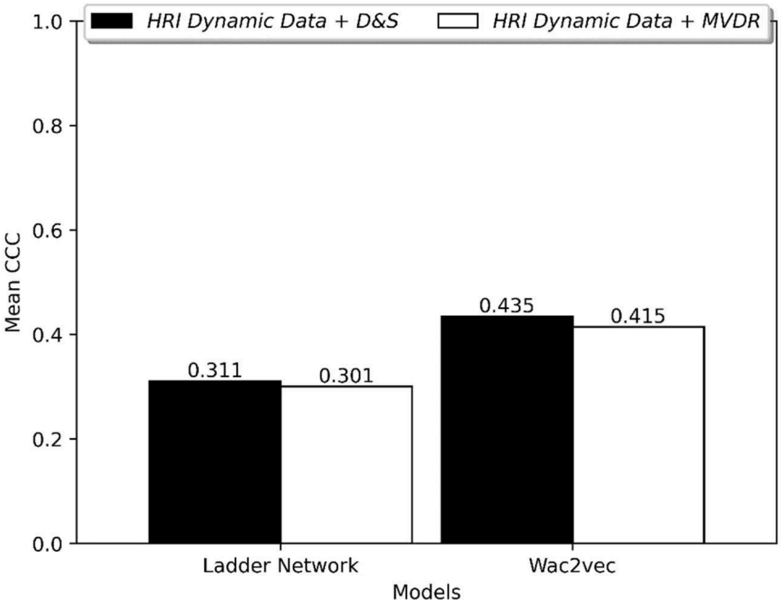

Fig. 10. Average CCC across valence, arousal, and dominance according to Table 5. The ladder network and wav2vec-based classifiers were trained with the MSP-Podcast training subset that had the acoustic model incorporated in the utterances according to Section 3.1. The testing subset corresponded to the testing partition of MSP-Podcast that was re-recorded in the dynamic condition with the HRI testbed according to Section 2.2.

2.0-based model takes better advantage of the simulated database condition because of the attentional scheme that models better the sequence of frames and its pre-training condition.

4.4. The static and dynamic HRI platform: a challenging testbed

The main contribution of this study is the acoustic channel framework proposed to mitigate the detrimental effect in SER performance expected in HRI where the distance and angle between the robot and the user vary. To the best knowledge of the authors, this is the first SER study in dynamic HRI, which is critical for practical applications where the distance and angle between the speakers and the microphone change during an interaction. While humans are quite good at understanding speech even in this dynamic scenario where the speech sources are moving, this scenario challenges most existing approaches that assume a static environment where the audio source is fixed. Our study investigates this relevant problem in the context of user profiling, where the robot aims to infer the emotional state of the users. We evaluate the proposed solutions using learned and handcrafted features, comparing the results in the presence of static and dynamic HRI scenarios.

Unlike the area of "robot profiling", which aims at making the robot capable of expressing human emotions through facial expressions or voice intonation (Stock-Homburg, 2022; Cameron et al., 2018; Ilić et al., 2019; Stock, 2019; Stock, 2016), the area of "user profiling" aims at making the robot capable of identifying the emotions of its human interlocutor (Rossi et al., 2017; Rajendran et al., 2023; Maroto-Gómez et al., 2023). The importance of SER development in HRI is evident (Sönmez and Varol, 2022): highlights the importance of improving the ability of robots to interact with humans more naturally and effectively. Understanding emotions holds a pivotal role in human communication, enabling robots to comprehend and respond to human needs more adeptly. In addition, it is mentioned that SER in HRI is a field that, even though virtually all social robots have microphones, has surprisingly been almost unexplored in the literature.

Most of the research on SER in HCI often overlooks the impact of the acoustic channel due to the proximity of users to the microphone (Shah Fahad et al., 2021; Alnuaim et al., 2022; Mustaqeem et al., 2023; Alnuaim et al., 2022; Mustaqeem and Kwon, 2021), unlike in HRI scenarios. Moreover, there are entire literature reviews on SER techniques that do not mention the application of these in HRI (Atmaja et al., 2022; Singh and Goel, 2022). As described in the introduction, none of the few studies addressing SER in HRI (Chen et al., 2020; Leem et al., 2021; Salekin et al., 2017; Ahmed et al., 2017) have considered the effects of a dynamic environment, i.e. with relative movement between the user and the robot. They have targeted only the effect of the human-robot distance.

There are two different features for SER: handcrafted and learned. On the one hand, Hashem et al. (2023) presents diverse handcrafted features applicable in both time and frequency domains, including energy, Frequency, Voice Quality, Spectral, and Cepstral features (Lee and Narayanan, 2005; Rao et al., 2013; Gao et al., 2017). On the other hand, the learned feature approach involves neural networks fed with voice signals to derive relevant representations for the task, predominantly applied in close-talk microphone HCI settings (Hashem et al., 2023). In this paper, we have managed to evaluate both approaches, both in static and dynamic environments. For the handcrafted features, we use the HLDs proposed in Schuller et al. (2013), implementing the approach with ladder networks. For the learned feature approach, we use wav2vec 2.0 (Wagner et al., 2023) which is fed with raw data. Thus, we compare the performance of both approaches under HRI conditions.

Regarding the direct comparison between wav2vec 2.0 and the ladder network, the former always provided a greater CCC. This must be because the wav2vec 2.0 is pre-trained with data that includes noisy conditions and extracts learned features. Fig 6 shows that wav2vec 2.0 always led to improvements in CCC when compared with the ladder network with and without beamforming techniques when both architectures were trained with the original MSP-Podcast training dataset. An interesting result is the comparison of D&S with MVDR. When combined with the ladder network, D&S did not provide a greater average CCC than MVDR. However, this situation was modified with wav2vec 2.0 and D&S usually gave greater average CCC than MVDR. As it is well known, speech enhancement techniques can introduce artifacts even though the SNR could be high (Goh et al., 1998) and MVDR is not an exception (Avila et al., 2016; Cauchi et al., 2019). If these artifacts can degrade the accuracy of ASR systems (Iwamoto et al., 2024), they can also deteriorate the performance of any speech-based pattern recognition technology including SER. Consequently, if wav2vec 2.0 provides greater CCC than the ladder network due to the pre-training conditions and the learned features, it cannot remove the artifact introduced by MVDR. Observe that the wav2vec 2.0 pre-training data does not include the beamforming response. The difference between the averaged CCCs across all the training and testing conditions with the ladder network for D&S and MVDR is -0.002. In contrast, the difference between the averaged CCCs across all the training and testing conditions with wav2vec 2.0 for D&S and MVDR is 0.0344.

We would like to emphasize that the proposed strategy is not just data augmentation that increases the training set size by incorporating, for instance, simple speed or amplitude perturbations on utterances to generate copies of the training data. Actually, the number of training utterances was not modified. Also, the proposed scheme is not multi-condition training either in the sense that it was not necessary to record training data in many indoor environments. In fact, we argue that to record training speech in several operating conditions would not be necessary. Instead, we model the indoor testing environment from the reverberation, noise, and beamforming response points of view to simulate the testing condition in the training data. The procedure employed and studied here in the framework of SER is interesting because it allows us to cope with the problem of complex real HRI scenarios with limited training data. The key factors involved in these conditions are: static and dynamic acoustic conditions with respect to noise and reverberation; the impact of increased interference due to noise and reverberation as the speaker-microphone distance increases beyond one meter; the response of beamforming schemes; and the TVAC effect that is observed in dynamic situations (Novoa et al., 2021). Our experiments with the HRI platform employed here show that the combination of all these factors can degrade the performance of SER systems dramatically. However, modeling the indoor testing environment from the reverberation, noise, and also beamforming response points of view to simulate the testing scenarios to be incorporated in the training data led to substantial increases in SER performance in the static and dynamic HRI conditions. Consequently, the strategy studied here softens the need to record training speech in several operating HRI situations. It is worth highlighting that our findings can be generalized to any indoor distant SER task.

5. Conclusions

In this paper, the problem of continuous SER was addressed for the first time in both real static and dynamic HRI scenarios. Two SER classifiers were evaluated, one implemented with the ladder network, and one implemented with the wav2vec 2.0 speech representation. Three training sets were considered: the original MSP-Podcast training utterances; and the MSP-Podcast training utterances that were processed to simulate the target acoustic environment plus the response of beamforming schemes such as D&S or MVDR. The acoustic environment was emulated as follows: first, RIRs were estimated in the testing room for the target speech and each one of both external noise sources by combining three robot positions with 21 different robot head angular positions, resulting in 63 RIRs for the target speech and for each one of the two noise sources resulting in 189 RIRs; second, the clean utterances and the two noise sources were convoluted with the corresponding RIRs, where speech utterances and noise samples shared the same positions of the robot and its head in each case; third, the speech and noise sources convoluted with the corresponding RIRs and the robot noise were added all together; and fourth, the beamforming responses were incorporated to the training noisy utterances. It is worth highlighting that, in the robotic testbed employed here, each robot head position determines a given DOA. Similarly, four testing data sets were employed: the original MSP-Podcast testing utterances; the MSP-Podcast testing utterances processed to simulate the testing environment as with the training data; and the MSP-Podcast testing dataset re-recorded in our HRI platform in static and dynamic conditions.

The results reported here show that the wav2vec 2.0 architecture provided a CCC metric that is on average 30.7 % higher than the ladder network-based engine using the original MSP-Podcast training utterances and the static and dynamic HRI testing subsets. This must be a result of the fact that wav2vec 2.0 is pre-trained with noisy data and employs learned features. Even though MVDR could lead to an average SNR improvement that is 1.73 dB higher than the basic D&S, both provided very similar average CCC improvement approximately equal to 116 % using the ladder network trained with the original MSP-Podcast training utterances and tested with the static and dynamic HRI data. However, when the wav2vec 2.0 classifier was employed in the same training/testing conditions, D&S led to an improvement in the average CCC that is slightly greater than the one achieved with MVDR but is much smaller than with the ladder network. Again, this must be because wav2vec 2.0 was pre-trained with noisy speech, so there is less room for improvement, and due to the MVDR artifacts. Surprisingly, in contrast to previous results with ASR, the static and dynamic HRI testing subsets resulted in a similar average CCC metric across all the training conditions. This must be a consequence of the SER engines employed here that determine emotional attributes on an utterance-by-utterance basis and of the periodicity of the translational/rotational robot movement. Simulating the acoustic environment and incorporating the beamforming response into the training dataset provided the highest average CCC scores with both the static and dynamic HRI subsets. These average scores are just 29 % and 22 % lower than those achieved with the original MSP-Podcast training/testing utterances with the ladder network and wav2vec 2.0 architectures, respectively. This result suggests that modeling the acoustic environment can be a convenient strategy to address the problem of SER in HRI

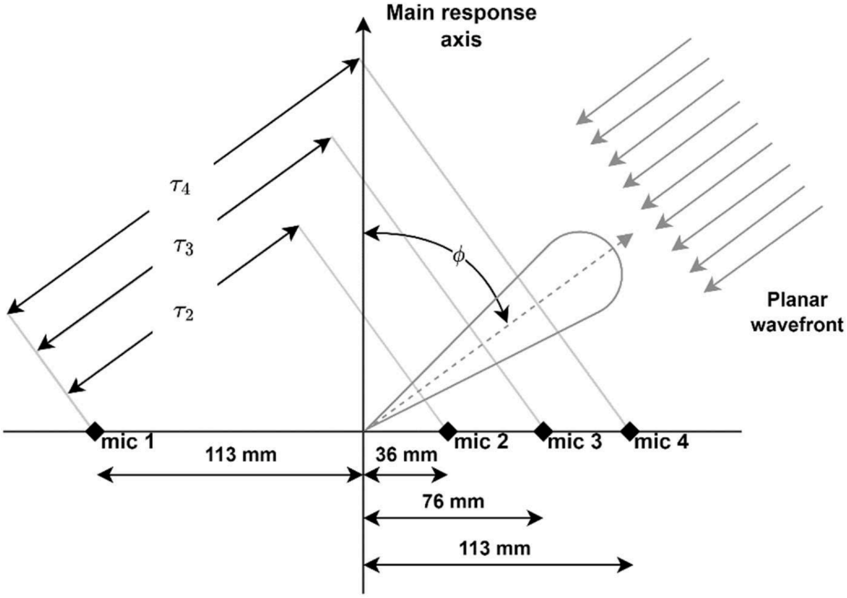

Fig. 11. Microphone array geometry of the Microsoft Kinect, where: τ_l is the time delay between microphone l and the reference one, i.e. microphone 1; and ϕ is the look direction or DOA.

scenarios without the need for recording ad-hoc training databases. Moreover, it is worth highlighting that our findings can be generalized to any indoor distant SER task. Finally, employing more complex robot movement conditions and integrating speech enhancement methods into the SER engines are proposed for future work.

CRediT authorship contribution statement

Nicolás Grágeda: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation. **Carlos Busso:** Writing – review & editing, Writing – original draft, Resources, Investigation, Conceptualization. **Eduardo Alvarado:** Software, Methodology, Data curation. **Ricardo García:** Writing – review & editing, Validation, Software. **Rodrigo Mahu:** Writing – review & editing, Writing – original draft, Software, Methodology. **Fernando Huenupan:** Validation, Data curation. **Néstor Becerra Yoma:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

The research reported here was funded by grant ANID (Chile)/FONDECYT 1211946.

Appendix

Beamforming techniques or spatial filtering denotes a family of technologies that is widely adopted to address distant speech processing. It has a very important role in social robotics, not only for speech-based HRI but also for analyzing audio sources. Two beamforming methods were considered here, D&S and MVDR.

A.1 Delay-and-sum

A microphone array is a collection of multiple microphones that can be combined and processed to achieve spatial filtering with beamforming. This technique can help to reduce noise and reverberation, especially by suppressing non-direct path acoustic signals. In this study, a linear microphone array was employed. Particularly, the Microsoft Kinect was adopted. This device is widely used in HRI applications, features a four-channel linear microphone array (see Fig. 11), and provides standard RGB and depth cameras.

Delay-and-sum is a well-known beamforming technique, which involves summing delayed signals to steer the look direction to the direction of arrival (DOA) of sound waves. This produces destructive interference in all directions except for DOA. The delayed signals are summed to generate the output signal y(t) as in Navvab et al. (2012):

$$y(t) = \sum_{l=1}^{L} x_l(t - \tau_l) \tag{1}$$

where: $x_l(t)$ denotes the signal samples from microphone l; τ_l is the delay applied to channel l with respect to the reference microphone, i.e. microphone 1 in this case; and L is the number of channels, i.e. L is equal to four here. Delay τ_l is given by Tashev (2009):

$$\tau_l = \frac{d_l}{v}(sen(\phi)) \tag{2}$$

where d_l is the distance between the microphone l and the reference one (see Fig. 11), and v is the propagation speed of sound.

A.2 MVDR

MVDR is a more advanced technique than delay-and-sum and improves beamforming noise suppression by adaptively reducing spatially correlated noise. This is achieved by creating nulls on the interfering signals without affecting the gain in the look direction. If $x_l(m,i)$ represents the i^{th}sample in frame m of channel l, where $1 \le m \le M$, M is the number of frames in a given utterance, $1 \le i \le$ *frameLength* and *frameLength* denotes the frame length in number of samples, $X_l(m,\omega)$ is obtained by applying the DFT to frame $x_l(m,i)$ and denotes the component at discrete frequency ω in frame m and channel l, where $0 \le \omega < numFreqBins$, $numFreqBins = DFT2 + 1$ and DFT corresponds to the number of samples employed by the DFT. The DFT of the MVDR output at the m^{th} frame, $Y(m,\omega)$, can be estimated as Kumatani et al. (2012):

$$Y(m,\omega) = w(m,\omega)[X_1(m,\omega)\ X_2(m,\omega)\ ...\ X_L(m,\omega)\ ] \tag{3}$$

Where the weights are estimated on a frame-by-frame basis as (Tashev, 2009):

$$w^H(m,\omega) = \frac{v^H(m,\omega)\sum_N^{-1}(m,\omega)}{v^H(m,\omega)\sum_N^{-1}(m,\omega)v(m,\omega)} \tag{4}$$

Eq. (4) includes the steering vector $v(m,\omega) = \left[e^{-j\omega\tau_1(m)}, e^{-j\omega\tau_2(m)}, ..., e^{-j\omega\tau_L(m)}\right]^T$ and the covariance matrix of the noise $\sum_N(m,\omega) = E\{N(m,\omega)N^H(m,\omega)\}$, where $E\{\cdot\}$ denotes the expectation operator and $N(\omega) = [N_1(\omega)N_2(\omega)...N_l(\omega)...N_L(\omega)]$ is the spatially correlated noise in each microphone in the frequency domain.

A.3 Delay-and-sum vs. MVDR

The primary focus of this paper is not only to compare the performance of the D&S and MVDR beamforming schemes. According to the strategy followed here, we argue that recording training speech in several operating HRI conditions would not be necessary. Instead, we model the indoor testing environment from the reverberation, noise, and also beamforming response points of view to simulate the testing condition in the training data. However, a brief comparison between D&S and MVDR can be interesting. First of all, D&S is the most basic beamforming scheme and requires only the DOA information to estimate and apply the resulting channel delays. In contrast, MVDR needs the estimation of the noise covariance matrix, which in turn could be a problem within speech intervals. This was counteracted by making the noise covariance matrices in speech segments equal to the interpolation of the matrices corresponding to the pre and post-noise intervals. It is worth highlighting that MVDR provides a larger SNR gain than D&S. Nevertheless, MVDR can incorporate artifacts (Avila et al., 2016; Cauchi et al., 2019) that degrade the accuracy of ASR systems (Iwamoto et al., 2024).

References

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

Alnuaim, A.A., et al., 2022a. Human-computer interaction for recognizing speech emotions using multilayer perceptron classifier. J. Healthc. Eng. https://doi.org/10.1155/2022/6005446.

Alnuaim, A.A., et al., 2022b. Human-computer interaction with detection of speaker emotions using convolution neural networks. Comput. Intell. Neurosci. https://doi.org/10.1155/2022/7463091.

Atmaja, B.T., Sasou, A., Akagi, M., 2022. Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion. Speech Commun. 140 https://doi.org/10.1016/j.specom.2022.03.002.

A. Avila, B. Cauchi, S. Goetze, S. Doclo and T. Falk. 'Performance comparison of intrusive and non-intrusive instrumental quality measures for enhanced speech'. 2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC), Xi'an, China, 2016, pp. 1–5, doi: 10.1109/IWAENC.2016.7602907.

Berg, J., Lottermoser, A., Richter, C., Reinhart, G., 2019. Human-robot-interaction for mobile industrial robot teams. Procedia CIRP. 79 https://doi.org/10.1016/j.procir.2019.02.080.

Bitzer, J., Simmer, K.U., 2001. Superdirective microphone arrays. In: Brandstein, M., Ward, D. (Eds.), Microphone Arrays: Signal Processing Techniques and Applications. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 19–38. https://doi.org/10.1007/978-3-662-04619-7_2. Eds.

Burmania, A., Parthasarathy, S., Busso, C., 2016. Increasing the reliability of crowdsourcing evaluations using online quality assessment. IEEE Trans. Affect. Comput. 7 (4) https://doi.org/10.1109/TAFFC.2015.2493525.

Busso, C., et al., 2008. IEMOCAP: interactive emotional dyadic motion capture database. Lang. Resour. Eval. 42 (4) https://doi.org/10.1007/s10579-008-9076-6.

C. Busso, M. Bulut, and S. Narayanan, 'Toward effective automatic recognition systems of emotion in speech', in Social Emotions in Nature and Artifact, J. Gratch and S. Marsella, Eds., Oxford University Press, 2013, pp. 110–127. doi: 10.1093/acprof:oso/9780195387643.003.0008.

Cameron, D., et al., 2018. The effects of robot facial emotional expressions and gender on child–robot interaction in a field study. Conn. Sci. 30 (4) https://doi.org/10.1080/09540091.2018.1454889.

Cao, H., Cooper, D.G., Keutmann, M.K., Gur, R.C., Nenkova, A., Verma, R., 2014. CREMA-D: crowd-sourced emotional multimodal actors dataset. IEEE Trans. Affect. Comput. 5 (4) https://doi.org/10.1109/TAFFC.2014.2336244.

B. Cauchi, K. Siedenburg, J.F. Santos, T.H. Falk, S. Doclo and S. Goetze. 'Non-intrusive speech quality prediction using modulation energies and LSTM-network.' in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 7, pp. 1151–1163, July 2019, doi: 10.1109/TASLP.2019.2912123.

Chakraborty, P., Ahmed, S., Yousuf, M.A., Azad, A., Alyami, S.A., Moni, M.A., 2021. A human-robot interaction system calculating visual focus of human's attention level. IEEE Access. 9 https://doi.org/10.1109/ACCESS.2021.3091642.

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

Deng, J., Pang, G., Zhang, Z., Pang, Z., Yang, H., Yang, G., 2019. CGAN based facial expression recognition for human-robot interaction. IEEE Access. 7 https://doi.org/10.1109/ACCESS.2019.2891668.

Devillers, L., Vidrascu, L., Lamel, L., 2005. Challenges in real-life emotion annotation and machine learning based detection. Neural Netw. 18 (4) https://doi.org/10.1016/j.neunet.2005.03.007.

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

Erdogan, H., Hershey, J., Watanabe, S., Mandel, M., Roux, J.Le, 2016. Improved MVDR beamforming using single-channel mask prediction networks. In: Proceedings of the Annual Conference of the International Speech Communication Association. INTERSPEECH, pp. 1981–1985. https://doi.org/10.21437/INTERSPEECH.2016-552 vol. 08-12-September-2016.

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.

A. Farina, 'Simultaneous measurement of impulse response and distortion with a swept-sine technique', Proc. AES 108th conv, Paris, France, no. I, 2000.

Y. Gao, B. Li, N. Wang, and T. Zhu, 'Speech emotion recognition using local and global features', in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2017. doi: 10.1007/978-3-319-70772-3_1.

A. Gaschler, S. Jentzsch, M. Giuliani, K. Huth, J. de Ruiter, 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.

Zenton Goh, Kah-Chye Tan and T.G. Tan. 'Postprocessing method for suppressing musical noise generated by spectral subtraction,' in IEEE Transactions on Speech and Audio Processing, vol. 6, no. 3, pp. 287–292, May 1998, doi: 10.1109/89.668822.

Grageda, N., Busso, C., Alvarado, E., Mahu, R., Becerra Yoma, N., 2023. Distant speech emotion recognition in an indoor human-robot interaction scenario. In: Proceedings of the Annual Conference of the International Speech Communication Association. INTERSPEECH.

Hashem, A., Arif, M., Alghamdi, M., 2023. Speech emotion recognition approaches: a systematic review. Speech Commun. 154 https://doi.org/10.1016/j.specom.2023.102974.

Hsu, W.N., et al., 2021. Robust wav2vec 2.0: analyzing domain shift in self-supervised pre-training. In: Proceedings of the Annual Conference of the International Speech Communication Association, 3. INTERSPEECH. https://doi.org/10.21437/Interspeech.2021-236.

J. Huang, Y. Li, J. Tao, Z. Lian, M. Niu and J. Yi, "Speech emotion recognition using semi-supervised learning with ladder networks", 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), Beijing, China, 2018, pp. 1–5, doi: 10.1109/ACIIAsia.2018.8470363.

D. Ilić, I. Žužić, and D. Bršćić, 'Calibrate my smile: robot learning its facial expressions through interactive play with humans', in HAI 2019 - Proceedings of the 7th International Conference on Human-Agent Interaction, 2019. doi: 10.1145/3349537.3351890.

Iwamoto, K., et al., 2024. How does end-to-end speech recognition training impact speech enhancement artifacts?. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, Korea. Republic of, pp. 11031–11035. https://doi.org/10.1109/ICASSP48485.2024.10447750.

Kousi, N., Stoubos, C., Gkournelos, C., Michalos, G., Makris, S., 2019. Enabling human robot interaction in flexible robotic assembly lines: an augmented reality based software suite. Procedia CIRP. 81 https://doi.org/10.1016/j.procir.2019.04.328.

Kumatani, K., et al., 2012. Microphone array processing for distant speech recognition: towards real-world deployment. In: 2012 Conference Handbook - Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. APSIPA ASC.

Lee, C.M., Narayanan, S.S., 2005. Toward detecting emotions in spoken dialogs. IEEE Trans. Speech Audio Process. 13 (2) https://doi.org/10.1109/TSA.2004.838534.

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.

W.-C. Lin and C. Busso. 'Chunk-level speech emotion recognition: a general framework of sequence-to-one dynamic temporal modeling'. in IEEE Transactions on Affective Computing, vol. 14, no. 2, pp. 1215–1227, 1 April-June 2023, doi: 10.1109/TAFFC.2021.3083821.

Lotfian, R., Busso, C., 2019. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. IEEe Trans. Affect. Comput. 10 (4) https://doi.org/10.1109/TAFFC.2017.2736999.

Maroto-Gómez, M., Marqués-Villaroya, S., Castillo, J.C., Castro-González, Á., Malfaz, M., 2023. Active learning based on computer vision and human–robot interaction for the user profiling and behavior personalization of an autonomous social robot. Eng. Appl. Artif. Intell. 117 https://doi.org/10.1016/j.engappai.2022.105631.

Metallinou, A., Yang, Z., chun Lee, C., Busso, C., Carnicke, S., Narayanan, S., 2016. The USC CreativeIT database of multimodal dyadic interactions: from speech and full body motion capture to continuous emotional annotations. Lang. Resour. Eval. 50 (3) https://doi.org/10.1007/s10579-015-9300-0.

Miseikis, J., et al., 2020. Lio-A personal robot assistant for human-robot interaction and care applications. IEEe Robot. Autom. Lett. 5 (4) https://doi.org/10.1109/LRA.2020.3007462.

E. Mower et al., 'Interpreting ambiguous emotional expressions', in Proceedings - 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII 2009. doi: 10.1109/ACII.2009.5349500. I. J. Tashev, Sound Capture and processing: Practical Approaches. John Wiley & Sons, 2009.

Mustaqeem, Kwon, S., 2021. MLT-DNet: speech emotion recognition using 1D dilated CNN based on multi-learning trick approach. Expert. Syst. Appl. 167 https://doi.org/10.1016/j.eswa.2020.114177.

Mustaqeem, K., El Saddik, A., Alotaibi, F.S., Pham, N.T., 2023. AAD-Net: advanced end-to-end signal processing system for human emotion detection & recognition using attention-based deep echo state network. Knowl. Based. Syst. 270 https://doi.org/10.1016/j.knosys.2023.110525.

M. Navvab, G. Heilmann, and A. Meyer, 'Simulation, visulization and perception of sound in a virtual environment using Beamforming', in Berlin, Beamforming Conference, Feb22-23, 2012.

J. Novoa, J. Wuth, J.P. Escudero, J. Fredes, R. Mahu, and N.B. Yoma, 'DNN-HMM based Automatic Speech Recognition for HRI Scenarios', in ACM/IEEE International Conference on Human-Robot Interaction, 2018. doi: 10.1145/3171221.3171280.

Novoa, J., Mahu, R., Wuth, J., Escudero, J.P., Fredes, J., Yoma, N.B., 2021. Automatic speech recognition for indoor HRI scenarios. ACM. Trans. Hum. Robot. Interact. 10 (2) https://doi.org/10.1145/3442629.

Omologo, M., Matassoni, M., Svaizer, P., 2001. Speech recognition with microphone arrays. In: Applications, M. (Ed.), Microphone Arrays: Signal Processing Techniques and. Springer Berlin Heidelberg, Heidelberg, pp. 331–353. https://doi.org/10.1007/978-3-662-04619-7_15. Brandstein and D. Ward, Eds. Berlin.

L. Paletta et al., 'Gaze-based human factors measurements for the evaluation of intuitive human-robot collaboration in real-time'. 2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), Zaragoza, Spain, 2019, pp. 1528–1531, doi: 10.1109/ETFA.2019.8869270.

Parthasarathy, S., Busso, C., 2018. 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. https://doi.org/10.21437/Interspeech.2018-1391.

Parthasarathy, S., Busso, C., 2020. Semi-supervised speech emotion recognition with ladder networks. IEEE/ACM. Trans. Audio Speech. Lang. Process. 28 https://doi.org/10.1109/TASLP.2020.3023632.

Rajendran, H., Bandara, H.M.R.T., Jayasekara, A.G.B.P., Chandima, D.P., 2023. User profiling based proactive interaction manager for adaptive human-robot interaction. 2023 Moratuwa Eng. Res. Confer. (MERCon), Moratuwa, Sri Lanka 632–637. https://doi.org/10.1109/MERCon60487.2023.10355527.

Rao, K.S., Koolagudi, S.G., Vempada, R.R., 2013. Emotion recognition from speech using global and local prosodic features. Int. J. Speech. Technol. 16 (2) https://doi.org/10.1007/s10772-012-9172-2.

Rossi, S., Ferland, F., Tapus, A., 2017. User profiling and behavioral adaptation for HRI: a survey. Pattern. Recognit. Lett. 99 https://doi.org/10.1016/j.patrec.2017.06.002.

Salekin, A., et al., 2017. Distant Emotion Recognition. Proc. ACM. Interact. Mob. Wearable Ubiquitous. Technol. 1 (3) https://doi.org/10.1145/3130961.

Scherer, K.R., 2003. Vocal communication of emotion: a review of research paradigms. Speech. Commun. 40 (1–2) https://doi.org/10.1016/S0167-6393(02)00084-5.

Schuller, B., et al., 2013. The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In: Proceedings of the Annual Conference of the International Speech Communication Association. INTERSPEECH. https://doi.org/10.21437/interspeech.2013-56.

Shah Fahad, M., Ranjan, A., Yadav, J., Deepak, A., 2021. A survey of speech emotion recognition in natural environment. Digi. Signal Process. Rev. J. 110 https://doi.org/10.1016/j.dsp.2020.102951.

K.U. Simmer, J. Bitzer, and C. Marro, 'Post-filtering techniques', 2001. doi: 10.1007/978-3-662-04619-7_3.

Singh, Y.B., Goel, S., 2022. A systematic literature review of speech emotion recognition approaches. Neurocomputing 492. https://doi.org/10.1016/j.neucom.2022.04.028.

Stock, R.M., 2016. Emotion transfer from frontline social robots to human customers during service encounters: testing an artificial emotional contagion modell. In: International Conference on Information Systems. ICIS, 2016 Proceedings.

Stock, R., et al., 2019. When robots enter our workplace: understanding employee trust in assistive robots. In: 40th International Conference on Information Systems. ICIS.

Stock-Homburg, R., 2022. Survey of emotions in human–robot interactions: perspectives from robotic psychology on 20 years of research. Int. J. Soc. Robot. 14 (2), 389–411. https://doi.org/10.1007/s12369-021-00778-6. May.

Y.Ü. Sönmez and A. Varol, 'The necessity of emotion recognition from speech signals for naturaland effective human-robot interaction in society 5.0′, in 2022 10th International Symposium on Digital Forensics and Security (ISDFS), pp. 1–8, 2022, doi:10.1109/ISDFS55398.2022.9800837.

J.H. Tao, J. Huang, Y. Li, Z. Lian, and M.Y. Niu, 'Semi-supervised ladder networks for speech emotion recognition', Int. J. Autom.Comput., vol. 16, no. 4, 2019, doi: 10.1007/s11633-019-1175-x.

Tashev, I.J., 2009. Sound Capture and processing: Practical Approaches. John Wiley & Sons.

Thiemann, J., Ito, N., Vincent, E., 2013. The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): a database of multichannel environmental noise recordings. Proc. Meetings Acoust. 19 (1), 035081 https://doi.org/10.1121/1.4799597. May.

J. Wagner et al., "Dawn of the transformer era in speech emotion recognition: closing the valence gap," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 9, pp. 10745–10759, 1 Sept. 2023, doi: 10.1109/TPAMI.2023.3263585.

Wang, Y., Boumadane, A., Heba, A., 2021. A fine-tuned Wav2vec 2.0/HuBERT benchmark for speech emotion recognition, speaker verification and spoken language understanding. CoRR. https://doi.org/10.48550/arXiv.2111.02735 vol. abs/2111.02735.

Yang, G.Z., et al., 2018. The grand challenges of science robotics. Sci. Robot. 3 (14) https://doi.org/10.1126/scirobotics.aar7650.