INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Respiratory distress estimation in human-robot interaction scenario

Eduardo Alvarado[1], *Nicolás Grágeda*[1], *Alejandro Luzanto*[1], *Rodrigo Mahu*[1], *Jorge Wuth*[1], *Laura Mendoza*[2], *Richard Stern*[3], *Néstor Becerra Yoma*[1]

[1]Speech Processing and Transmission Laboratory, University of Chile
[2]Clinical Hospital, University of Chile
[3]Department of Electrical and Computer Engineering, Carnegie Mellon University
nbecerra@ing.uchile.cl, eduardo.alvarado@ug.uchile.cl

Abstract

Social robotics and human-robot partnership are becoming very relevant topics in the next decades defining many challenges for speech technology. In addition, the COVID pandemic imposed an awareness of technology challenges to fight massive health problems. In this paper, the first system to estimate respiratory distress in a human-robot interaction (HRI) environment is presented. The training procedure of the dyspnea estimation models by simulating the HRI acoustic environment with real room impulse responses (estimated with a PR2 robot) and additive noise is described. The training and testing data were processed using two beamforming techniques: delay-and-sum and MVDR. The results suggest that it should be possible to reduce significantly the degradation in precision of estimates of respiratory distress in a real HRI scenario. The improvements in accuracy and AUC with MVDR when compared to baseline processing without beamforming are 7% and 4%, respectively.

Index Terms: Respiratory distress estimation, human-robot interaction, speech based user profiling.

1. Introduction

1.1. Social Robots and user profiling

Social robots are designed to communicate and coordinate with people to achieve common goals. These types of robots are especially potentially useful in areas such as education or healthcare [1]. To emulate human communication successfully, it is necessary for the robot to characterize the user's profile physically, cognitively and/or socially [2]. In this way, the robot can adapt its response based on the user's behavior and needs. Physically characterizing the person is a complicated task, as obtaining the information needed to do so is often very invasive, such as blood pressure measurements, blood tests or lung capacity measurements. Other less invasive options involve the use of wearables, which allow measurements of sleep, movement, neurology, cardiovascular status, etc., in a fast and comfortable way for the user [3]. However, wearables are not accurate or robust, so improving their performance remains a major challenge [4], [5].

In this context, the use of voice emerges as an important alternative for user profiling in HRI [6]. The voice includes linguistic and paralinguistic information (prosody) which are especially useful in several applications [7], and it also allows the capture of information to detect respiratory problems [8].

1.2. Estimation of Respiratory Distress

In recent years, the growing demand for health services has led to a significant increase in the development of remote health monitoring tools [9]. Significant efforts have been made to prevent her COPD using machine learning (ML). Because these methods are effective in collecting and integrating large-scale disparate medical data in precision medicine [10]. This development has been driven by the COVID-19 pandemic [11] and has resulted in the development of several artificial intelligence (AI)-based solutions for automatically detecting SARS-CoV-2 [12]. These ML-based solutions were primarily focused on smartphones due to the great scalability, ubiquity, and flexibility offered by these devices [13].

Various studies, methods and databases focused on remote monitoring of respiratory diseases by voice analysis was performed in the state-of-art [14]. Detection is done by voice analysis, so evaluation can be standardized, resulting in reduced variability or bias between different physicians who administer the test in the form of questionnaires. Moreover, the responses given in a questionnaire can be influenced by the patient's mood or habituation to the disease [15]–[17] .

1.3. Human-robot interaction in respiratory distress

Although the confinements caused by the pandemic seem to have come to an end, problems in healthcare centers persist, such as lack of supplies, shortage of professionals and the growth of vulnerable populations. This is where an ideal environment arises for social robots to intrude into this world. Surprisingly, the use of social robots to estimate dyspnea is null, especially considering the rise of the independent respiratory distress and Human-robot interaction (HRI) studies in recent years.

As mentioned above, several studies postulate the use of voice for estimating respiratory disorders. However, the effects of external conditions such as additive noise (noisy cocktail parties), reverberation and/or movement of the speakers has not been addressed in the literature. In contrast, classical beamforming schemes such as delay-and-sum (D&S) [18] and MVDR [19] have widely been used in the literature to filter the target signal spatially [20].

In this paper, an automatic dyspnea detection system is proposed in an HRI environment. This design allows monitoring the degree of respiratory distress on the modified Medical Research Council (mMRC) scale that classifies dyspnea into five levels, from zero (healthy) to four (very severe). Surprisingly, this topic has not been studied, since there

10.21437/Interspeech.2023-963

are no studies that combine the evaluation of respiratory distress with HRI.

In [21] a system to estimate dyspnea on the phone with deep learning was proposed. The database that was used to train the system was composed of three types of controlled vocalizations that were obtained by prompting the users to take deep breaths and to vocalize them until gasping for air. The idea was to represent the level of dyspnea by characterizing the user's behavior while pronouncing the controlled vocalizations. The first two phonetizations correspond to /ae-ae/and/sa-sa/ [21]. They provide relevant information about the amount of air exhaled by the individuals. In contrast to sustained vowels employed elsewhere, they are not stationary and are not cancelled by the noise suppression schemes in smart phones. The third phonetization corresponded to counting from one to thirty as fast as possible, to assess the spontaneous behavior of the subjects who must make an effort to reach the goal. The idea was to cause involuntary breathing, voice pauses, coughing, tone variation, etc., that could characterize dyspnea severity [21].

The method described in [21] extracted time-dependent and time-independent features from each vocalization. The time-independent features were processed by MLP based classifiers. In contrast, the time-dependent features employed CNN based architectures for /ae-ae/ and /sa-sa/ vocalizations, and a CNN-LSTM based for the one-to-30 counting.

In this paper, we train dyspnea estimation models by simulating the HRI acoustic environment with real room impulse responses (RIR) and additive noise as suggested in [22] for speech recognition. The RIRs were estimated with our robotic platform composed of a PR2 robot and studio loudspeakers. Additionally, two beamforming algorithms were evaluated: D&S and MVDR. The results reported here suggests that it should be possible to reduce significantly the precision degradation of the respiratory distress estimation in a HRI scenario. The main contributions of this paper are: design and evaluation of a respiratory distress estimation system in HRI, which is the first one to our knowledge; and, the incorporation of acoustic modelling and beamforming methods to improve the robustness of this system in HRI scenarios.

2. HRI Scenario Testing Databases

2.1. Database

The database employed here was generated from the one used in [21] after asking again the informed consent from all subjects. Those that agreed to participate in this study were included in the new database. This is composed of patients recruited at the Clinical Hospital at the University of Chile (HCUCH) and of healthy volunteers from the Faculty of Physical and Mathematical Sciences (FCFM) at the same university. The scientific ethics committees at HCUCH and FCFM approved the study.

Individuals were prompted to produce the three controlled vocalizations without pauses after taking deep breaths until they gasped for air. The first vocalization is the Spanish phoneme sequence /a/ and /e/, denoted as /ae-ae/; the next one is the Spanish syllable sequence /sa/, denoted as /sa-sa/; and the last one was inspired by the Roth Test [23] where subjects were asked to count in Spanish from one to 30 as fast as they could.

The database had 100 people, of whom 66 were patients with respiratory problems persons (39 COPD, 22 Pulmonary Fibrosis and 5 sequelae of COVID-19) and 34 were healthy. Clinical evaluation of the patients revealed that 18 had mMRC scores of one, 27 had mMRC scores of two, 19 had mMRC scores of three, and two had mMRC scores of four. The healthy participants received an mMRC mark equal to zero. Our deep learning-based models were trained using the clinical evaluation (Gold standard) as references. Only two patients presented mMRC scores equal to four and they were incorporated in the subset corresponding to mMRC score equal to three. Hence, four mMRC levels were considered from zero to three.

Each type of vocalization was repeated twice by each individual. As a result, the whole database was composed of two repetitions per type of phonetization and per individual x three types of phonetizations x 100 individuals = 600 vocalizations. An automatic speech recognition (ASR) system was trained to separate the target vocalizations from the background noise or unwanted audio after recording.

2.2. Testing and training database

The original telephone database mentioned above was used to generate training data by simulating the acoustic model of the HRI scenario as in [22]. All the 600 audios that compose the database were convolved with 33 real RIRs obtained with the testbed shown in Fig. 1 in 33 static positions: three robot-audio speaker distances denoted by positions P1, P2 and P3; and 11 angles of the robot head. Figure 2 shows the experimental robotic setup where the RIRs were estimated. In addition, additive noise was added at SNRs within the range of 5 dB and 15 dB. The additive noise was generated by combining the PR2 engine noise with noise samples from the Aurora database (street, cars, babble, airport, train, subway) at SNRs within the range of -5 dB to 5 dB. The Aurora noise signals at each microphone had the phases modified to simulate different DOAs (direction of arrival) and represent real HRI scenarios as close as possible.

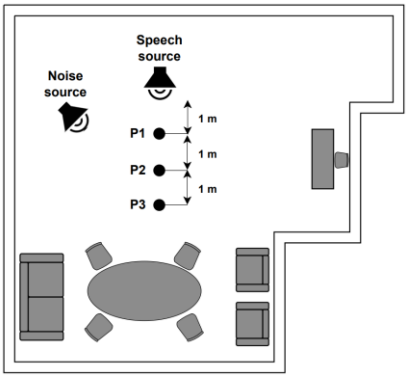


Figure 1: *Plan view of room used for RIRs estimation and HRI simulated scenario.*

Figure 2: *The experimental setup used to record RIRs.*

3. Respiratory distress estimation system in HRI

The block diagram of the proposed system is shown in Figure 3. First, it is assumed that acoustic sources can be localized with sensors (e.g. cameras) in the robot providing the DOA information for the beamforming technology. This is particularly important in indoor environments where DOA estimation is affected by both reverberation and noise sources. Second, the enhanced target source signals and the acoustic model of the HRI indoor scenario are employed to train and estimate the mMRC score.

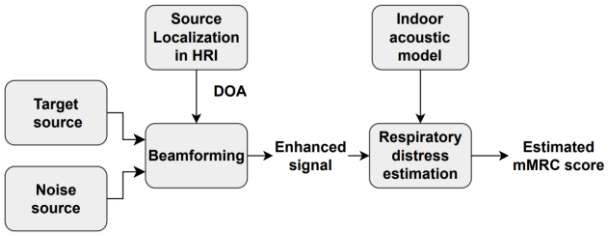


Figure 3: *Block diagram of the respiratory distress estimation in HRI scenario.*

3.1. Source localization

The PR2 robot automatically saves the azimuthal angle of its head, which provides the DOA info required by the beamforming schemes. This is considered an input from the beamformer point of view.

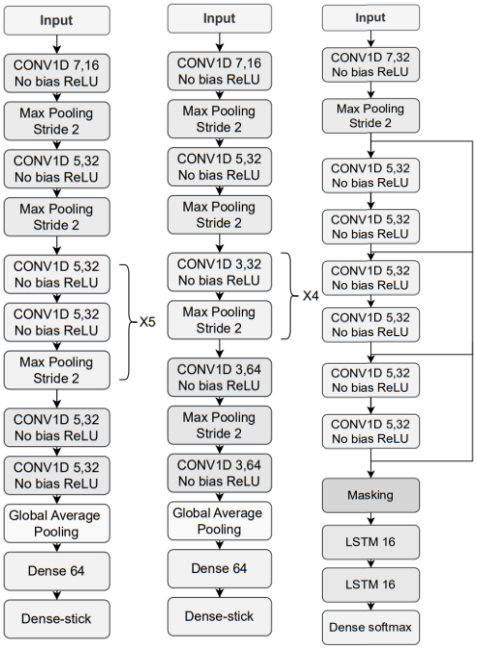
3.2. Beamforming based speech enhancement

The second block of the proposed system consists of speech enhancement. Here the aim is to improve the quality of the target speech signal by employing spatial filtering. In this paper, two beamforming methods were tested: D&S and MVDR.

3.3. Deep learning based respiratory distress estimation module

The system's goal is to classify users' dyspnea levels on the mMRC scale based on how they behave while performing controlled vocalizations. This system based on telephone speech is described in detail in [21], and is based on characterizing the users' spontaneous behavior while vocalizing controlled phonetizations. The time-dependent features are computed frame-by-frame and attempt to capture the dynamics of the voice signals. On the other hand, time-independent features provide information concerning the whole vocalizations.

The MLP and CNN/LSTM architectures and training configurations in [21] were optimized again in this paper without additive noise, where the optimal configuration was obtained when the validation-1 subset (see Section 3.4) returned the best accuracy and AUC. For time-independent features, the MLP network had two hidden layers with 40 neurons each and a learning rate equal to 0.01 in the case of /ae-ae/. In the MLP corresponding to /sa-sa/, the optimal setup corresponding to two hidden layers with 20 nodes each and a learning rate of 0.01 was adopted. Finally, five hidden layers of 10 nodes each were found to be optimal in the case of the one-to-30 counting, with a learning rate equal to 0.001. On the other hand, the time-dependent features are based on the FFT log power spectrum and were optimized for each type of phonetization. 512-sample FFTs are estimated in 50-ms windows with 50% overlap where 257 frequency bins are obtained. After that, 14 logarithmic Mel-frame filter energies are calculated for phonetizations /ae-ae/ and /sa-sa/. For the one-to-30 counting, 75% of the lowest frequency bins were chosen to compute the log spectrum. Then, the corresponding first derivative or delta features were included only in one-to-30 counting. Parameter means and variances are calculated for the entire database, and MVN is applied to the temporal trajectories of the time-dependent features. The time-dependent feature architecture and hyper parameter optimization resulted in the use of neuron stick breaking, cross-entropy as a loss function, the use of the ADAM optimizer, and a learning rate of 0.0005 for /ae-ae/ and /sa-sa/, and of 0.0001 for the one-to-30 counting. Figure 4 depicts the resulting deep learning architectures for the three types of vocalizations.



(a) /ae-ae/; (b) /sa-sa/; (c) *one-to-30 counting*

Figure 4: *Neural network architectures for time dependent features.*

3.4. K-fold training with double validation

To optimize the available database, a nine-fold cross-validation was performed, from which 11 users were removed from each of the partitions for testing, except for one partition from where 12 individuals were removed. It is important to mention that this data division scheme ensures that a given speaker could not have vocalizations in the training, validation, or testing subsets simultaneously. Besides the testing individuals, each partition was composed of training, validation-1 and validation-2 subsets corresponding to 70%, 15% and 15% of the partition individuals, respectively. The classifiers were trained and tested as in [21].

The classifiers were trained eight times with each partition. Subset validation-1 data was employed to stop the iterations with an early stopping of 20. For each partition, the optimal model was chosen by picking the one with the highest average accuracy on the validation-1 and validation-2 subsets. The

whole procedure was repeated five times to obtain more reliable statistics. A GeForce GTX 1080 GPU was employed.

3.5. Acoustic modeling training for respiratory distress estimation in HRI

As mentioned above, the proposed acoustic model was based on convolving the original audio signals with real RIRs estimated at static positions and adding additive noise [22]. Moreover, noise was added emulating different DOAs and the responses of the D&S and MVDR beamforming methods were also incorporated in the model.

4. Results

In this section we show the performance of the respiratory distress system when performing matched training and testing databases. As in our work described in [21], the results correspond to propagating the test subsets of the nine folds (Section 3.4) through the corresponding trained models. It is worth highlighting that the nine-fold testing subsets cover the whole 100 subject database.

Figures 5 and 6 show accuracy (i.e. the percentage of trials for which the estimated and reference mMRC scores are the same) and the area under the curve (AUC), respectively, with the original telephone (*Telephone*), simulated HRI with no additive noise (*Sim-HRI-noiseless*), simulated HRI with additive noise (*Sim-HRI-noisy*), HRI with additive noise and D&S response (*Sim-HRI-D&S*) and HRI with additive noise and MVDR response (*Sim-HRI-MVDR*). As seen in Figs. 5 and 6, the model trained and tested with *Sim-HRI-noiseless* provided an accuracy and AUC with combined features that is 11% and 0.5% lower, respectively than those obtained with the model trained and tested with the original conditions employed in [21], i.e. *Telephone*. This must be due to the effect of the distortion introduced by reverberation. As expected, the worst performance was observed with *Sim-HRI-noisy*, which in turn provided an accuracy and AUC degradation with combined features equal to 5% and 5%, respectively, when compared with *Sim-HRI-noiseless*. These results of further distortion introduced by additive noise. In contrast, the incorporation of the beamforming response in the training and testing process, i,e, *Sim-HRI-D&S* and *Sim-HRI-MVDR*, provided much lower average degradations in AUC (3%) with combined features than *Sim-HRI-noisy* when compared with *Sim-HRI-noiseless*, even in accuracy improves by 3%. This result strongly suggests that spatial filtering can be of a great help in distant respiratory distress assessment even in matched training and testing conditions. It is interesting to observe that MVDR gave slightly worse accuracy than D&S. This might be a result of the artifacts introduced by this beamforming scheme.

As can be seen in Figs. 5 and 6, in most cases the best performance was achieved with the combination of time-dependent and time-independent based classifiers. These results confirm the pertinence of the definition and combination scheme of the engineered features employed here. Another critical point to highlight is the greater robustness of the time-independent parameters. These coefficients led to standard deviations in accuracy and AUC of 1.52% and 0.021%, respectively, across all the five training/testing conditions. In contrast, the standard deviations in accuracy and AUC obtained with time-dependent parameters were 3.15% and 0.027%, respectively, across all the five training and testing scenarios. This result suggests that, despite the fact that the time-dependent features provide more information than the time-independent ones, the latter depend less on the training/testing scenarios than the former.

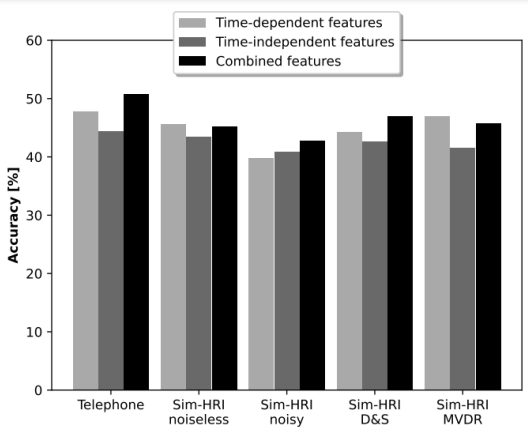


Figure 5: *Accuracy in simulated data for time dependent, time independent and combined features.*

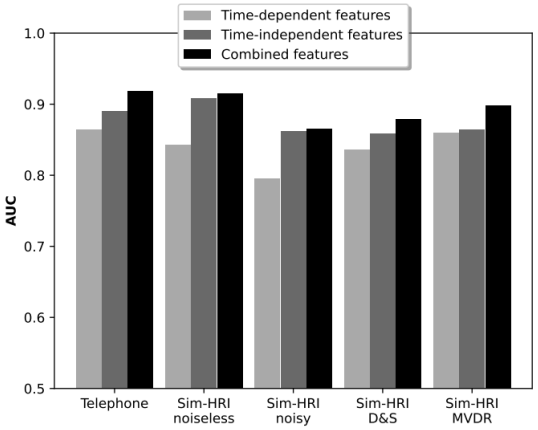


Figure 6: *AUC in simulated data for time dependent, time independent and combined features.*

5. Conclusions

In this paper we present the first system to estimate respiratory distress in a human-robot interaction environment. A telephone database was processed to simulate an HRI scenario. We proposed to train the dyspnea estimation models by simulating the HRI acoustic environment with real room impulsive responses (estimated with a PR2 robot and loud speakers) and additive noise, similar to our previous work with a speech recognition task in previous papers. Additionally, the training and testing data were also processed with the response of two beamforming techniques, i.e. delay-and-sum and MVDR. The results suggest that it should be possible to reduce significantly the degradation in precision of estimates of respiratory distress in a real HRI scenario. The improvements in accuracy and AUC with MVDR when compared with the baseline situation without spatial filtering are 7% and 4%, respectively. Addressing the problem of respiratory distress estimation in real static and dynamic HRI conditions is proposed as future research.

6. Acknowledgements

This research was funded by ANID/FONDECYT grant No. 1211946 and ANID/COVID grant No. 0365.

7. References

[1] C. Breazeal, K. Dautenhahn, and T. Kanda, "Social Robotics," in *Springer Handbook of Robotics*, B. Siciliano and O. Khatib, Eds., Cham: Springer International Publishing, 2016, pp. 1935–1972. doi: 10.1007/978-3-319-32552-1_72.

[2] S. Rossi, F. Ferland, and A. Tapus, "User profiling and behavioral adaptation for HRI: A survey," *Pattern Recognit Lett*, vol. 99, pp. 3–12, 2017, doi: https://doi.org/10.1016/j.patrec.2017.06.002.

[3] J. Dunn, R. Runge, and M. Snyder, "Wearables and the medical revolution," *Per Med*, vol. 15, no. 5, pp. 429–448, 2018.

[4] J. Tana, M. Forss, and T. Hellsten, "The use of wearables in healthcare–challenges and opportunities," 2017.

[5] M. Smuck, C. A. Odonkor, J. K. Wilt, N. Schmidt, and M. A. Swiernik, "The emerging clinical role of wearables: factors for successful implementation in healthcare," *NPJ Digit Med*, vol. 4, no. 1, pp. 1–8, 2021.

[6] J. Wuth, P. Correa, T. Núñez, M. Saavedra, and N. Yoma, "The Role of Speech Technology in User Perception and Context Acquisition in HRI," *Int J Soc Robot*, vol. 13, Feb. 2021, doi: 10.1007/s12369-020-00682-5.

[7] J. Cole, "Prosody in context: A review," *Lang Cogn Neurosci*, vol. 30, no. 1–2, pp. 1–31, 2015.

[8] K. K. Lella and A. PJA, "A literature review on COVID-19 disease diagnosis from respiratory sound data," *arXiv preprint arXiv:2112.07670*, 2021.

[9] R. Duggal, I. Brindle, and J. Bagenal, "Digital healthcare: Regulating the revolution," *BMJ (Online)*, vol. 360. 2018. doi: 10.1136/bmj.k6.

[10] Y. Feng, Y. Wang, C. Zeng, and H. Mao, "Artificial intelligence and machine learning in chronic airway diseases: Focus on asthma and chronic obstructive pulmonary disease," *International Journal of Medical Sciences*, vol. 18, no. 13. 2021. doi: 10.7150/ijms.58191.

[11] J. F. Veenis, S. P. Radhoe, P. Hooijmans, and J. J. Brugts, "Remote monitoring in chronic heart failure patients: Is non-invasive remote monitoring the way to go?," *Sensors (Switzerland)*, vol. 21, no. 3. 2021. doi: 10.3390/s21030887.

[12] N. Subramanian, O. Elharrouss, S. Al-Maadeed, and M. Chowdhury, "A review of deep learning-based detection methods for COVID-19," *Computers in Biology and Medicine*, vol. 143. 2022. doi: 10.1016/j.compbiomed.2022.105233.

[13] S. Amiriparian and B. Schuller, "AI Hears Your Health: Computer Audition for Health Monitoring," in *Communications in Computer and Information Science*, 2021. doi: 10.1007/978-3-030-94209-0_20.

[14] T. Xia, J. Han, and C. Mascolo, "Exploring machine learning for audio-based respiratory condition screening: A concise review of databases, methods, and open issues," *Exp Biol Med*, vol. 247, no. 22, pp. 2053–2061, 2022.

[15] M. C. Stoeckel, R. W. Esser, M. Gamer, C. Büchel, and A. von Leupoldt, "Brain mechanisms of short-term habituation and sensitization toward dyspnea," *Front Psychol*, vol. 6, no. JUN, 2015, doi: 10.3389/fpsyg.2015.00748.

[16] L. Wan, L. Stans, K. Bogaerts, M. Decramer, and O. van den Bergh, "Sensitization in medically unexplained dyspnea: Differential effects on intensity and unpleasantness," *Chest*, vol. 141, no. 4, 2012, doi: 10.1378/chest.11-1423.

[17] A. von Leupoldt and B. Dahme, "Psychological aspects in the perception of dyspnea in obstructive pulmonary diseases," *Respiratory Medicine*, vol. 101, no. 3. 2007. doi: 10.1016/j.rmed.2006.06.011.

[18] R. Zahn, J. D. Johnston, and G. W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *Journal of the Acoustical Society of America*, vol. 78, no. 5, 1985, doi: 10.1121/1.392786.

[19] Y. Xiao, J. Yin, H. Qi, H. Yin, and G. Hua, "MVDR algorithm based on estimated diagonal loading for beamforming," *Math Probl Eng*, vol. 2017, 2017.

[20] N. Saleem and M. I. Khattak, "A review of supervised learning algorithms for single channel speech enhancement," *Int J Speech Technol*, vol. 22, no. 4, pp. 1051–1075, 2019.

[21] E. Alvarado, N. Grageda, A. Luzanto, R. Mahu, J. With, L. Mendoza and N. B. Yoma, "Dyspnea Severity Assessment Based on Vocalization Behavior with Deep Learning on the Telephone," *Sensors*, vol. 23, no. 5, 2023, doi: 10.3390/s23052441.

[22] J. Novoa, R. Mahu, J. Wuth, J. P. Escudero, J. Fredes, and N. B. Yoma, "Automatic speech recognition for indoor hri scenarios," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 10, no. 2, pp. 1–30, 2021.

[23] E. Chorin, A. Padegimas, O. Havakuk, E. Birati, Y. Shacham, A. Milman, G. Topaz, N. Flint, G. Keren and O. Rogowski, "Assessment of Respiratory Distress by the Roth Score," *Clin Cardiol*, vol. 39, no. 11, 2016, doi: 10.1002/clc.22586.