INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

RaSSpeR: Radar-based Silent Speech Recognition

David Ferreira[1,2], *Samuel Silva*[1,2], *Francisco Curado*[1,2], *António Teixeira*[1,2]

[1]Dep. Electronics Telec. & Informatics, University of Aveiro, Aveiro, Portugal
[2]Institute of Electronics and Informatics Engineering of Aveiro (IEETA), Portugal
davidcruzferreira@ua.pt, sss@ua.pt, fcurado@ua.pt, ajst@ua.pt

Abstract

Speech is our most natural and efficient way of communication and offers a strong potential to improve how we interact with machines. However, speech communication can sometimes be limited by environmental (e.g., ambient noise), contextual (e.g., need for privacy in a public place), or health conditions (e.g., laryngectomy), hindering the consideration of audible speech. In this regard, silent speech interfaces (SSI) have been proposed (e.g., considering video, electromyography), however, many technologies still face limitations regarding their everyday use, e.g., the need to place equipment in contact with the speaker (e.g., electrodes/ultrasound probe), and raise technical (e.g., lighting conditions for video) or privacy concerns. In this context, the consideration of technologies that can help tackle these issues, e.g, by being contactless and/or placed in the environment, can foster the widespread use of SSI. In this article, continuous-wave radar is explored to assess its potential for SSI. To this end, a corpus of 13 words was acquired, for 3 speakers, and different classifiers were tested on the resulting data. The best results, obtained using Bagging classifier, trained for each speaker, with 5-fold cross-validation, yielded an average accuracy of 0.826, an encouraging result that establishes promising grounds for further exploration of this technology for silent speech recognition.

Index Terms: continuous-wave radar, silent speech recognition, European Portuguese, machine learning

1. Introduction

Speech is a natural and efficient form of human communication, and, as such, the research on speech technologies that can foster its use in domains such as Human-Computer Interaction (HCI) is highly relevant. While Automatic Speech Recognition (ASR) is commonly used in HCI environments, as in the case of Amazon's Alexa and Apple's Siri, there are still some scenarios that cannot take the most out of speech interaction, including situations where privacy is needed, environmental noise is present, silence is required, or in the most extreme cases, when health conditions incapacitate speakers to produce acoustic signals.

To tackle such scenarios, Silent Speech Interfaces (SSI) emerged as a possible alternative to consider, consisting of the process of speech communication in the absence of an audible/intelligible acoustic signal. Since speech production is a complex motor process, which starts in the brain and ends with respiratory, laryngeal, and articulatory motion, each step of its production process can be explored and physiologically measured through specialized sensors and methods to potentially infer what the speaker is trying to say without relying on the acoustic signal.

Although there are already a lot of proposed sensors and technologies for silent speech recognition, most of them have characteristics that limit their use in everyday life, e.g., by being intrusive, non-portable, affected by noise, user-dependent, or just not affordable [1, 2, 3, 4].

In light of these challenges, it is important to explore novel and improved technologies that might bring SSI to a wider variety of scenarios and users. In this regard, SSI technologies that could be placed in the environment, instead of on the person, and potentially used by any passing user, might open new scenarios of application. To this end, frequency-modulated continuous wave (FMCW) radar technology, already used in a wide variety of scenarios including the automotive industry and service robots, emerges as a possible candidate to tackle some of these issues given its noninvasiveness, nonintrusiveness, portability, and independence of ambient lighting. Furthermore, recent evolutions have made it less costly and easily available commercially, making radar technology appear in many of our daily environments. The recent market launch of the first mobile phone with radar (Google pixel 4; 2019) dedicated to proximity detection and identification of manual commands from the user without direct contact with the device, points towards the vulgarization of this technology and opens up prospects of its future exploitation for novel applications. In this regard, the exploration of a technology that might already be present to bring SSI capabilities to the environment is a promising path to follow. However, while a strong potential can be anticipated, given these perceived advantages, radar technology has yet to prove its mettle for silent speech applications. To assess the extent of its applicability, this work performs a first exploration of FMCW radar technology and shows its viability for silent speech recognition.

The remainder of this document is structured as follows: Section 2 presents a brief overview on related work regarding noninvasive SSI, also covering previous research on SSIs for European Portuguese; section 3 describes the adopted methods, from environment and acquisition settings to data exploration, feature extraction, and classification approaches; section 4 reports the classification results, obtained for the collected dataset, and in section 5 these are further analyzed and discussed; finally, section 6 presents some concluding remarks and ideas for further advancing this work.

2. Related work on SSI

A distinguishing element of SSIs is speech recognition beyond the acoustic signal, exploring other biosignals associated with the different stages of the speech production process. From brain waves to the visual aspects of speech, several approaches have been developed to allow silent speech recognition (SSR). While relevant work also exists for invasive technologies, the overview provided in what follows privileges noninvasive methods, in line with our goals.

In SSI development research, surface EMG (sEMG) is the most common technology used, since it is easy to apply and is

http://dx.doi.org/10.21437/Interspeech.2021-1413

less prone to raise ethical concerns for volunteers [5, 6]. Recent work has privileged the evolution and consideration of increasingly imperceptible and highly flexible sEMG electrodes (e.g., [7, 8, 9, 10, 11]) and notable results include those of Liu et al. [7] and Dong et al. [6], where accuracies greater than 80% (6 words) and 70% (3 words) were achieved. Nevertheless, there is still a high data variability between sessions due to the participants' skin impedance [12].

Non-audible murmur (NAM) microphones are another technology widely used in SSI development to capture and record murmured speech and other smooth vocal productions resultant from the acoustic output. Recognition rates of nearly 70% were achieved in a corpus containing 421 total utterances [13]. However, NAM is highly user-dependent due to participants' physiological differences [13].

Electroencephalography (EEG) enables measuring electrical brain signals in a non-invasive way. Recent studies (e.g., [14, 15, 16]) present accuracy rates ranging the 70% mark for corpora of 5 syllables, 7 phrases, and 5 phrases, respectively. One clear advantage of this technology is that it allows visualizing the activation of the different brain areas associated with speech production, but it is highly sensitive to noise and its recognition rates are user-dependent.

Ultrasound (US) imaging is another technology widely considered in SSI research, as it allows observing tongue movement sequences during the speech production process. Some recent works include those of Chen et al. [17] and Xu et al. [18], in which, respectively, a new technique for representing speech articulation resorting to an ultrasound-driven finite element model of the tongue is presented, and a novel sequential feature extraction approach for SSI systems is explored. Nevertheless, while US images yield good spatial and temporal resolutions, the images have relatively low quality due to the presence of speckle noise [19].

Video imaging can be used to capture visible speech articulators. By resorting to different models and algorithms, which allow extracting the articulators of interest from each frame, it is possible to obtain accuracy rates as high as 91% for a corpus of 22 syllables and 95% for one of 44 phrases (e.g., [20, 21, 22, 23, 24]). Although generally being low-cost, this technology is susceptible to raise privacy concerns and is strongly affected by ambient illumination.

Regarding contactless radar-based silent speech recognition, and to the best of our knowledge, not much has yet been explored apart from a recent work by Shin et al. [4]. In this study, a recognition rate of 85% was reported for a corpus comprising 10 isolated words considering a dynamic time warping (DTW) approach. However, as stated by the authors, some limitations resided in the fact that distance and correlation amplitude were the only considered features, and that there was recognition degradation due to slight head movements of the participants throughout the acquisition sessions, something that would require additional methods to mitigate.

Concerning SSI for European Portuguese (EP), several technologies have been researched (e.g., EEG, sEMG, ultrasonic doppler (UD), Video, and Depth), also including multimodal approaches [3]. Freitas et al. [25] proposed Visual Speech Recognition (VSR) and Acoustic Doppler Sensors (ADS) for silent speech recognition, resorting to Dynamic Time Warping (DTW), achieving an 8.6% Word Error Rate (WER). Later, in 2013, the same author [26] selected 4 non-invasive modalities (Visual data from Video and Depth, sEMG and UD) and proposed a system that explores their synchronous combination into a multimodal SSI. The same corpus as the one explored in this article was considered, and DTW and KNN were used. It was verified that the combination of multiple modalities presented a better performance (93.8%), while subsets of the modalities produced lower results, such as 71.4% for the Video and Depth combination. In more recent work, Teixeira et al. [27], proposed an approach based on VSR to enable real-time control of a media player, having achieved an accuracy of 81.3% for a corpus comprising 8 control commands.

3. Methods

As the main goal of this work was to perform a first assessment of FMCW radar-based technology regarding its silent speech recognition capabilities, our main purpose was to establish an approach that was a compromise between keeping some aspects closer to a real scenario, e.g., no chin rest during acquisitions, but establishing some controlled conditions, such as head orientation towards the radar. This, without compromising our main purpose, should reduce some of the complexity of the data acquisition and postprocessing, at this stage.

3.1. Radar Equipment and Speaker-to-Radar Positioning

The board considered for this investigation was the AWR1642BOOST-EVM from Texas Instruments, an evaluation board for the AWR1642 FMCW radar sensor. This board is currently used at our research institute for multiple purposes ranging from robot navigation and human detection [28] to biosignal measurement [29]. The room selected for acquisition was free from any moving objects other than the participant, ensuring no interference in the acquired data.

Considering the orientation of the participants' heads towards the radar, acquisitions with several orientations would further inform about the capabilities of the technology in everyday scenarios. However, considering the added complexity to the data acquisitions, and given the main overall purpose of this work, i.e., a first viability assessment of the technology for SSR, data was only acquired while speakers faced the radar.

Finally, the distance at which the participants would be from the radar was another aspect taken into account. Since the proximity of the participant towards the radar and the consequent occupation of a large portion of its field of view would yield more data points, at this stage, we established the distance between the speaker and the radar at about 15cm, as if it was a microphone. Nevertheless, and since we are aiming for reduced intrusiveness, we opted for not fixing the position of the participants' head: they were instructed to try and keep the same relative position to the radar, but without enforcing it. However, it is worth noting that, as something that will be explored in future work, by exploiting the MIMO technology and beamforming capabilities of this radar board, the field-of-view (FoV) of the sensor can be reduced, allowing the user to stay at a larger distance from the radar while allowing for the acquisition of reflection data focused in a comparable area of the body.

3.2. Data Acquisition

For the data acquisition sessions, we resorted to Texas Instruments' DemoVisualizer application, a software that enables radar configuration, data acquisition, and data visualization. DemoVisualizer enabled testing different radar configurations while focusing on the acquisition aspects that most suit the particular research experience. As previously explained, we privileged a less fixed head position, and this, as observed by [4], might yield added challenges in using distance to the radar as

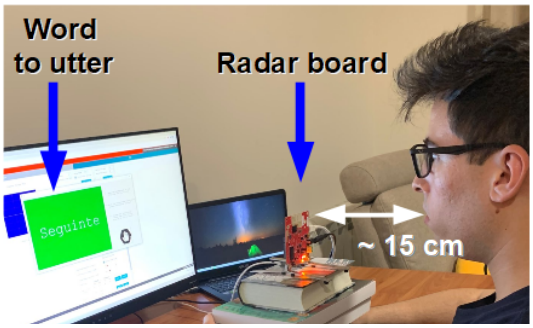

Figure 1: *Radar setup disposition. The participant is seated in front of the radar board while a monitor continuously displays the words to be uttered, turning the background green whenever the participant is asked to speak.*

the data considered for recognition. Therefore, we opted for a configuration that prioritized the best possible velocity resolution envisaging having the participants' facial velocity dispersion while they produced speech.

Custom software was developed to automatically manage the data acquisition procedure. The participants were placed in front of the radar board, at approximately 15 cm from it (Figure 1). An LCD display, placed in the participant's line of sight, provided information about the word to be uttered. After informed consent, the speakers were only asked to speak at a normal rhythm and the acquisition procedure started. For each trial, a random word of the corpus would appear on the display, and a beep (along with a change in color of the screen) would signal that the speaker could utter it. After the beep, the acquisition software recorded radar data for two seconds.

3.3. Corpus

Considering the team's work history on Ambient Assisted Living (AAL) and its previous work in SSI research for these contexts, we adopted a previously considered corpus [26, 30]. Therefore, thirteen EP words were considered:

Ajuda	(help)	**Anterior**	(previous)
Calendário	(calendar)	**Contactos**	(contacts)
Email	(email)	**Familia**	(family)
Fotografias	(photographs)	**Lembretes**	(reminders)
Ligar	(turn on)	**Mensagens**	(messages)
Pesquisar	(search)	**Seguinte**	(next)
Vídeos	(videos)		

Three participants enrolled in the data acquisition sessions: (a) one of the authors, an Engineering Ph.D. student, 26 years old, male; (b) a 24 years old female Psychologist; and (c) a 50 years old female, a real estate manager. All participants were native EP speakers. The inclusion of one of the authors as a participant aimed to gather data from a speaker that, while still speaking in a normal way, would pay attention to keeping a more consistent speaking pattern throughout the acquisition session. This would inform a best-case scenario where a prospective user would be instructed to be consistent in uttering the words.

For each of the 13 words present in the corpus, 60 validated repetitions were considered per participant. The validation stage mainly ensured the removal of the recordings in which no usable data was produced (e.g., participant missing the recording slot or data recording error).

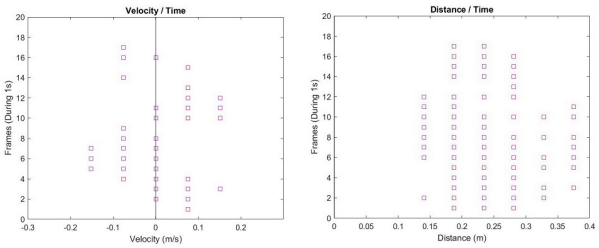

Figure 2: *Illustrative visual representations for the data corresponding to one acquisition of the word "Ajuda" (Help): velocity dispersion pattern (left) and distance variation (right) over 1 second of acquisition frames (along the vertical axis). For this case study, although distance variations were acquired and presented, only the velocity representations were considered for model training and subsequent classification.*

3.4. Feature Extraction and Classification

After the acquisition sessions, it was necessary to explore the data and define a set of features to be tested and used for classification. During data acquisition, the signals received by the RF front-end of the board are digitized and pre-processed by the built-in ADC and DSP, respectively, and the raw data thus obtained is assembled into several tag-length-value (TLV) packets. To process these packets, a parser was developed in Matlab to extract all the detected objects' relevant information (i.e., their Cartesian coordinates (X, Y, Z) and relative velocities expressed in the radar frame of reference). These data includes the full point cloud detected in the radar FoV, over the time acquisition windows, from which distinct subsets of points are clustered and associated, in real-time, by the firmware of the board, to represent distinct objects, or parts of a body, with distinctive velocity measures. These include both static and moving objects and can also include "fake" objects, resulting from multiple reflections in walls or other surfaces. Therefore, a first step to enable the ensuing analysis was the elimination, from the data, of all static objects or objects that were more than 30 cm away from the radar board. Finally, for each word instance, two visual representations were created, one depicting distance variations over time, and another depicting the velocities dispersion over time (Figure 2).

Although distance information could potentially provide pertinent data for classification purposes, it was confirmed that small distance differences between recordings of the same word would produce substantially different representations. This was expected, as the participants could slightly move their heads throughout the acquisition sessions. While additional postprocessing could help minimize this issue, as mentioned in a previous work that mainly explored distance data [4], it was left for future work, and we ultimately opted for exploring the dispersion of velocity data associated with the users' facial motions.

3.4.1. Classifier Training and Testing

For testing the different classification approaches, the velocity dispersion data (as depicted in figure 2) from each word instance (words being the classes) is provided to the machine learning algorithms. The whole 2 s of recorded data per word instance were always considered, regardless of word duration. Several classifiers were considered for testing, namely, Random Forests (RF), Linear Discriminant Analysis (LDA), Linear Regression (LD), Support Vector Machine (SVM), and Bagging (BAG), to understand the impact they would have on the classification re-

Table 1: *Mean, standard deviation, and maximum value of accuracy obtained for different classifiers while considering the results for three speakers. The three rightmost columns present the mean accuracy values per speaker and classifier. All results were obtained considering 5-fold cross validation.*

	M	SD	max	Spk1	Spk2	Spk3
RF	0.816	0.072	0.923	0.881	**0.760**	0.806
BAG	**0.826**	0.092	**0.955**	**0.915**	0.740	**0.822**
LDA	0.783	0.095	0.929	0.887	0.717	0.745
LR	0.803	0.100	**0.955**	**0.915**	0.744	0.750
SVM	0.773	0.099	0.910	0.888	0.708	0.722

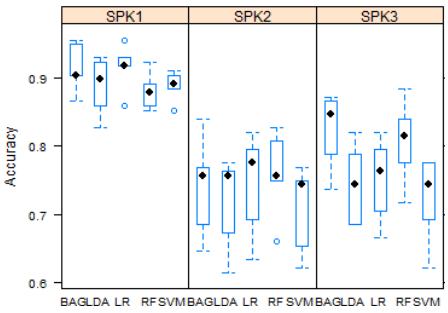

Figure 3: *Boxplot representation of all classifiers' accuracies for each speaker.*

sults for the acquired data. A 5-fold cross-validation approach was additionally considered due to the size of the data, ensuring that every observation from the original dataset had the chance of appearing in both training and testing. Accuracy was adopted for assessing the performance of the different classifiers given that the considered data set is class-balanced (i.e., no disparity between the number of instances belonging to each class).

4. Results

The summary of the obtained results, for the different classifiers, is depicted in Table 1. All mean accuracy values across all participants, obtained from the different classifiers, were greater than 0.75. The best mean accuracy was obtained from the BAG classifier ($M = 0.826; SD = 0.092$), having, as well, produced the maximum accuracy of 0.955 for a specific k-fold iteration. The classifier producing the lowest mean accuracy was SVM ($M = 0.773; SD = 0.099$), also showing the lowest maximum for a specific k-fold iteration (0.910).

Regarding the accuracy values obtained for each speaker and classifier, it is possible to verify that, for Speaker 1, both BAG and LR classifiers achieved the highest mean accuracy values (0.915), for Speaker 2, the RF classifier presented the highest mean accuracy value (0.760), and, for Speaker 3, the BAG classifier produced the highest mean accuracy value (0.822).

Considering the average accuracy results from all classifiers, for the different speakers, as depicted in Figure 3, it is possible to verify that Speaker 1 presented higher accuracy values than the other speakers which, in turn, present similar results.

Concerning the recognition rates for each of the words of the corpus, by analyzing the resulting confusion matrices, it was verified that the results were trendily worse for the words "Lembretes" (Reminders) and "Ligar" (Turn on) (with a correspondingly average accuracy, across all speakers, of 0.63 and 0.71), typically being erroneously classified as "Email" and "Seguinte" (Next), respectively.

5. Discussion

The main focus of this project was to assess radar-based technology as a viable solution for silent speech recognition. From the obtained results, with the average accuracy being as high as 82.6% for the BAG classifier, considering all speakers, it is a good indication of the capabilities of this technology, particularly since it yielded high accuracy rates for a set of 13 words. Overall, it is possible to observe that, independently of the classifier, Speaker 1 had the highest average accuracy. This is probably because this speaker was aware of the importance of producing the words consistently throughout the acquisition session. This rationale is supported by the boxplots depicted in Figure 3. While for Speaker 1 the boxes are tighter for all classifiers, for both Speaker 2 and 3 the boxes indicate more dispersion of the 5-fold cross-validation results, probably due to slight variations in the speech articulation throughout the acquisition sessions. This hints that such a simple instruction has the potential to improve performance. However, even for Speakers 2 and 3, that had no previous experience with radar technology and received no other indication than to speak normally, results as high as 76% and 82.2% were obtained.

Comparing the obtained recognition results with previous works for the same AAL corpus, radar-based SSI obtained either comparable or superior accuracy. In [30] an accuracy of 75% is reported for sEMG, and in [26] accuracies of 71.4%, 72.6%, and 83% were obtained for Video, Depth, and UDS technologies, respectively.

Regarding the lower recognition accuracies (lowest = 63%) for the words "l**em**bretes" (reminders), sometimes recognized as "**em**ail", and "li**g**ar" (turn on), sometimes recognized as "se**g**uinte" (next), this may be due to some notable articulatory similarity, e.g., at the beginning or middle of the word that, at some elocution speeds – and, eventually, as a consequence of coarticulatory effects –, may turn their velocity dispersion patterns more similar.

6. Conclusions

In this paper, we propose the consideration of an AWR1642 radar board to assess the viability of FMCW radar technology for SSR. Based on velocity dispersion features, several classification models were trained and were capable of producing results as good as 82.6% across all participants.

These results establish promising grounds for further exploring this technology for SSI, and future work will tackle aspects such as the possibility of creating speaker-independent systems, the influence of speaker-to-radar distance, or the improvements enabled by other classifiers (e.g., ANN, CNN).

7. Acknowledgements

Work partially funded by IEETA RU funding (UIDB/00127/2020), by Portugal 2020 under the Competitiveness and Internationalization Operational Program, and the European Reg. Dev. Fund through proj. MEMNON (POCI-01-0145-FEDER-028976) and proj. RETIOT – Reflectometry Technologies to Enhance the Future Internet of Things and Cyber-Physical Systems (POCI-01-0145-FEDER-016432, FCT-SAICTPAC/0006/2015). A word of thanks is due to the participants for their contributions.

8. References

[1] S. Ahmed and S. H. Cho, "Hand gesture recognition using an ir-uwb radar with an inception module-based classifier," *Sensors*, vol. 20, no. 2, p. 564, 2020.

[2] S. Hazra and A. Santra, "Short-range radar-based gesture recognition system using 3d cnn with triplet loss," *IEEE Access*, vol. 7, pp. 125 623–125 633, 2019.

[3] J. Freitas, A. Teixeira, M. S. Dias, S. Silva *et al.*, *An Introduction to Silent Speech Interfaces*. Springer, 2017.

[4] Y. H. Shin and J. Seo, "Towards contactless silent speech recognition based on detection of active and visible articulators using ir-uwb radar," *Sensors*, vol. 16, no. 11, p. 1812, 2016.

[5] G. S. Meltzner, J. T. Heaton, Y. Deng, G. De Luca, S. H. Roy, and J. C. Kline, "Development of semg sensors and algorithms for silent speech recognition," *Journal of neural engineering*, vol. 15, no. 4, p. 046031, 2018.

[6] W. Dong, H. Zhang, H. Liu, T. Chen, and L. Sun, "A superflexible and high-sensitive epidermal semg electrode patch for silent speech recognition," in *2019 IEEE 32nd International Conference on Micro Electro Mechanical Systems (MEMS)*. IEEE, 2019, pp. 565–568.

[7] H. Liu, W. Dong, Y. Li, F. Li, J. Geng, M. Zhu, T. Chen, H. Zhang, L. Sun, and C. Lee, "An epidermal semg tattoo-like patch as a new human–machine interface for patients with loss of voice," *Microsystems & Nanoengineering*, vol. 6, no. 1, pp. 1–13, 2020.

[8] A. F. Ruiz-Olaya and A. López-Delis, "Surface emg signal analysis based on the empirical mode decomposition for human-robot interaction," in *Symposium of Signals, Images and Artificial Vision-2013: STSIVA-2013*. IEEE, 2013, pp. 1–4.

[9] L. Diener, T. Umesh, and T. Schultz, "Improving fundamental frequency generation in emg-to-speech conversion using a quantization approach," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 682–689.

[10] J. E. Joy, H. A. Yadukrishnan, V. Poojith, and J. Prathap, "Work-in-progress: Silent speech recognition interface for the differently abled," in *International Conference on Remote Engineering and Virtual Instrumentation*. Springer, 2019, pp. 805–813.

[11] A. Kapur, S. Kapur, and P. Maes, "Alterego: A personalized wearable silent speech interface," in *23rd International conference on intelligent user interfaces*, 2018, pp. 43–53.

[12] R. Merletti and P. J. Parker, *Electromyography: physiology, engineering, and non-invasive applications*. John Wiley & Sons, 2004, vol. 11.

[13] N. Shah, N. J. Shah, and H. A. Patil, "Effectiveness of generative adversarial network for non-audible murmur-to-whisper speech conversion." in *INTERSPEECH*, 2018, pp. 3157–3161.

[14] L. Sarmiento, J. B. Rodríguez, O. López, S. Villamizar, R. Guevara, and C. Cortes-Rodriguez, "Recognition of silent speech syllables for brain-computer interfaces," in *2019 IEEE International Conference on E-health Networking, Application & Services (HealthCom)*. IEEE, 2019, pp. 1–5.

[15] D. Cao, D. Zhang, and H. Chen, "A novel task-oriented text corpus in silent speech recognition and its natural language generation construction method," in *Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval*, 2019, pp. 30–33.

[16] D. Dash, A. Wisler, P. Ferrari, and J. Wang, "Towards a speaker independent speech-bci using speaker adaptation." in *INTERSPEECH*, 2019, pp. 864–868.

[17] S. Chen, Y. Zheng, C. Wu, G. Sheng, P. Roussel, and B. Denby, "Direct, near real time animation of a 3d tongue model using non-invasive ultrasound images," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4994–4998.

[18] K. Xu, Y. Wu, and Z. Gao, "Ultrasound-based silent speech interface using sequential convolutional auto-encoder," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 2194–2195.

[19] T. Schultz, M. Wand, T. Hueber, D. J. Krusienski, C. Herff, and J. S. Brumberg, "Biosignal-based spoken communication: A survey," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2257–2271, 2017.

[20] T. Thein and K. M. San, "Lip localization technique towards an automatic lip reading approach for myanmar consonants recognition," in *2018 International Conference on Information and Computer Technologies (ICICT)*. IEEE, 2018, pp. 123–127.

[21] A. P. Kandagal, V. Udayashankara, and M. Anusuya, "Silent speech recognition," in *International Conference on Cognitive Computing and Information Processing*. Springer, 2017, pp. 130–139.

[22] K. Sun, C. Yu, W. Shi, L. Liu, and Y. Shi, "Lip-interact: Improving mobile device interaction with silent speech commands," in *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, 2018, pp. 581–593.

[23] Z. Thabet, A. Nabih, K. Azmi, Y. Samy, G. Khoriba, and M. Elshehaly, "Lipreading using a comparative machine learning approach," in *2018 First International Workshop on Deep and Representation Learning (IWDRL)*. IEEE, 2018, pp. 19–25.

[24] S. Uttam, Y. Kumar, D. Sahrawat, M. Aggarwal, R. R. Shah, D. Mahata, and A. Stent, "Hush-hush speak: Speech reconstruction using silent videos." in *INTERSPEECH*, 2019, pp. 136–140.

[25] J. Freitas, A. Teixeira, C. Bastos, and M. Dias, "Towards a multimodal silent speech interface for european portuguese," in *Speech technologies*. InTech, 2011, pp. 125–149.

[26] J. Freitas, A. Teixeira, and M. S. Dias, "Multimodal silent speech interface based on video, depth, surface electromyography and ultrasonic doppler: Data collection and first recognition results," in *Int. Workshop on Speech Production in Automatic Speech Recognition*, 2013.

[27] A. Teixeira, N. Vitor, J. Freitas, and S. Silva, "Silent speech interaction for ambient assisted living scenarios," in *International Conference on Human Aspects of IT for the Aged Population*. Springer, 2017, pp. 369–387.

[28] D. F. Albuquerque, E. S. Gonçalves, E. F. Pedrosa, F. C. Teixeira, and J. N. Vieira, "Robot self position based on asynchronous millimetre wave radar interference," in *2019 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. IEEE, 2019, pp. 1–6.

[29] C. Gouveia, A. Tomé, F. Barros, S. C. Soares, J. Vieira, and P. Pinho, "Study on the usage feasibility of continuous-wave radar for emotion recognition," *Biomedical Signal Processing and Control*, vol. 58, p. 101835, 2020.

[30] J. Freitas, "Articulation in multimodal silent speech interface for european portuguese," Ph.D. dissertation, University of Aveiro, 2015, [Online; accessed 28-march-2020]. [Online]. Available: https://ria.ua.pt/bitstream/10773/14425/1/Tese.pdf