Interspeech 2022
18-22 September 2022, Incheon, Korea

A user-friendly headset for radar-based silent speech recognition

Pouriya Amini Digehsara[1], *João Vítor Possamai de Menezes*[1], *Christoph Wagner*[1], *Michael Bärhold*[2], *Petr Schaffer*[2], *Dirk Plettemeier*[2], *Peter Birkholz*[1]

[1]Institute of Acoustics and Speech Communication, Technische Universität Dresden, Germany
[2]Institute of Communication Technology, Technische Universität Dresden, Germany

Pouriya.amini@tu-dresden.de

Abstract

Silent speech interfaces allow speech communication to take place in the absence of the acoustic speech signal. Radar-based sensing with radio antennas on the speakers' face can be used as a non-invasive modality to measure speech articulation in such applications. One of the major challenges with this approach is the variability between different sessions, mainly due to the repositioning of the antennas on the face of the speaker. In order to reduce the impact of this influencing factor, we developed a wearable headset that can be 3D-printed with flexible materials and weighs only about 69 g. For evaluation, a radar-based word recognition experiment was performed, where five speakers recorded a speech corpus in multiple sessions, alternatively with the headset and with double-sided tape to place the antennas on the face. By using a bidirectional long short-term memory network for classification, an average inter-session word accuracy of 76.50% and 68.18% was obtained using the headset and the tape, respectively. This indicates that the antenna (re-) positioning accuracy with the headset is not worse than that with the double-sided tape while providing other benefits.

Index Terms: silent speech interfaces, wearable headset, BiLSTM, radar imaging, speech-related biosignals

1. Introduction

Silent speech interfaces (SSIs) are systems that enable speech communication without any acoustical information [1]. This is possible because SSIs sense speech-related biosignals, e.g., muscle signals [2, 3], brain activity [4, 5], or articulatory movements [6, 7], in real-time during speech. Possible application scenarios of SSIs are where people are physiologically incapable of producing audible speech, where confidentiality is needed in public spaces, or where the acoustic noise level of the environment masks the audible speech [8]. For the broad use of these systems, they should be stable (make reproducible measurements), portable (as small as possible), convenient (easy to use) and non-invasive (without any sensors under the skin or inside the mouth). To address the stability issue, many researchers use mechanical systems to enable reproducible measurements, and the most common solution to date is the use of stabilizing headsets [9].

Multiple types of wearable systems are used in a variety of SSIs. In a surface electromyography-based (sEMG) system, a gypsum mask was used to improve the stability of the position of the sensors [3]. In magnetic sensors-based systems, sensors have been mounted directly on wearable systems such as glasses [10] and custom interfaces, e.g., a combination of tongue magnet and outer ear interfaces (TMI+OEI) [11]. In an ultrasound-based (US) system, a headset was developed to stabilize the US transducer and to reduce the discomfort of the subjects during recordings [12]. These wearable systems have, however, considerable shortcomings. The gypsum mask for sEMG sensors presented in [3] lacks portability and it is not comfortable for the users. While the measurements made with the glasses presented in [10] lack stability, the TMI+OEI system presented in [11] is considered as invasive, as magnets are attached to the tongue. While achieving good stability for the measurements, the headset presented in [12] has a high weight of around 350 g, which is uncomfortable for the subjects during longer recordings. The widespread usage of a wearable system depends on the simultaneous fulfilment of the stability, portability, convenience and non-invasiveness requirements, which the mentioned systems cannot fully meet yet.

Besides the aforementioned ones, radar-based SSIs are a promising alternative that was first proposed by Holzrichter et al. in 1998 [13]. Recent examples of such systems performed speech recognition experiments based on contactless monostatic radar to measure the reflection coefficient of the vocal tract [14] and on contactless bistatic radar to detect vocal tract movements [15]. Birkholz et al. [8] proposed another approach to radar-based SSIs with three antennas attached directly to the facial skin of the speaker (with medical-grade skin adhesive tape) and a standard network analyzer [16, 17]. The system showed high potential for distinguishing German phonemes in intra-session experiments.

In this study, we propose a novel portable headset for stable inter-session data acquisition based on the approach presented in [8, 17], and compare its performance with the one obtained with the tape-based fixation method of the antennas previously used.

2. Development of the headset

The goal was to create a mobile headset that allows a consistent positioning of the two required antennas on the left and right cheek of the speaker's face (one sending and one receiving antenna) while being as lightweight as possible and adaptable to various head shapes. All iterations of the mechanical frames were designed with Autodesk Inventor and manufactured using a 3D printer (Ultimaker 3), to enable accessability of the headset. To limit the weight of the headset, lightweight materials were used: flexible black thermoplastic polyurethane (TPU-95A), rigid grey polylactic acid (PLA) filament, elastic textile, and plastic bolts and nuts, as shown in Figure 1. To make the headset adaptable to different head shapes, the tightness of the headband and the chin band are adjustable. The position of the antennas on the face is determined by three variables (distances d_1, d_2 and the angle α, see Figure 2) that must be adjusted once for each speaker. Once adjusted, the antennas always come to rest on the same part of the face when the headset is put on again by the same speaker.

10.21437/Interspeech.2022-10090

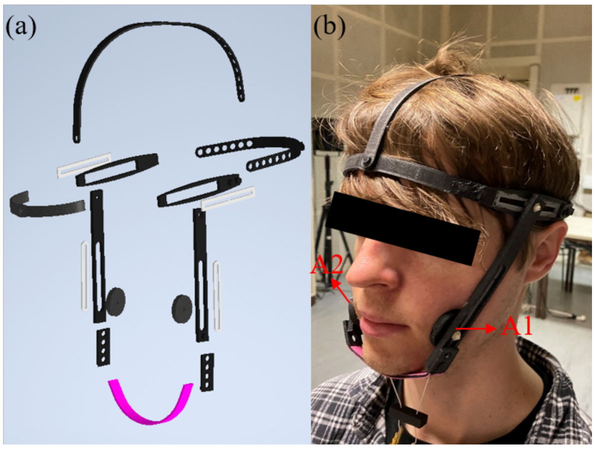

Figure 1: *The proposed headset. (a) Exploded view (black: TPU-95A, grey: PLA, purple: elastic textile). (b) Photograph of a subject wearing the headset. The two antennas are labeled as A1 and A2.*

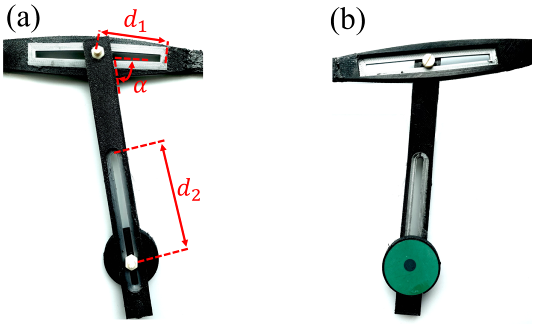

Figure 2: *Adjustable variables to adjust the antenna positions. (a) Back side. (b) Front side.*

All 3D-printable parts are available in the supplemental material at https://www.vocaltractlab.de/index.php?page=birkholz-supplements.

3. Methods

3.1. Recording hardware

The hardware used to record data in this study is different from the standard network analyzer used in [8]. Since a minimum measurement update rate of 100 Hz is necessary to capture articulatory movements in real-time, we developed a stepped-frequency continuous wave (SFCW) radar hardware capable of delivering measurements at this rate, shown in Figure 3. This hardware emits broadband EM signals from one port (TX) and simultaneously measures the signals received at the other port (RX). In this study, one antenna was connected to each port and attached to the speaker's cheeks, so the received signals are associated with the vocal tract shapes and hence with the speech sounds, as shown in [8]. The TX port was connected to antenna A1 in Fig. 1.

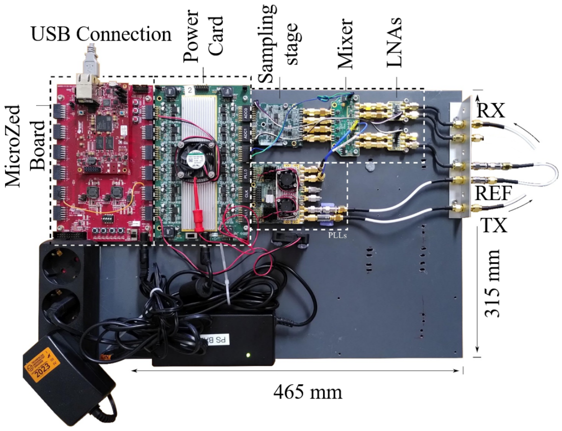

Figure 3: *Stepped-frequency continuous-wave radar hardware [17].*

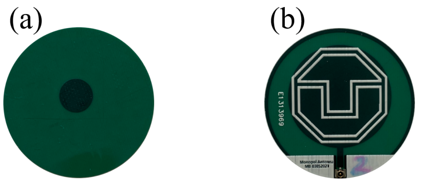

Figure 4: *Monopole antenna; (a) front side, (b) back side.*

The signal bandwidth for this study was changed from the one used in [8], which was 2-12 GHz. Since the received signals were strongly damped above 6 GHz with a correspondingly low signal-to-noise ratio, we decided to lower both band limits here and used a frequency band from 1 to 6 GHz. This also allows the exploration of signals from 1 to 2 GHz that were not used in [8].

To emit the frequencies over the full bandwidth (128 discrete frequencies linearly spaced along the band), the hardware uses two frequency synthesizers (type LMX2594, Texas Instruments) with an offset of 1 MHz between them. The emitted signal flows through port TX to antenna A1. The signals are then filtered by the vocal tract, received by antenna A2 and transmitted to the hardware through the port RX [17].

As antennas, we used printed circular monopole antennas with a diameter of 30 mm, as shown in Figure 4. Coplanar waveguides stimulate the antennas, flowing into a 20 mm diameter disc. This antenna was designed to emit EM waves from 1 to 6 GHz, so its dimensions are appropriate to resonate at these frequencies. The antennas were also designed to emit EM waves directly through the skin, which has a high permittivity. To this end, the chosen antenna characteristic type was a monopole, whose radiation consists of only one lobe, directed perpendicular into the skin of the face.

3.2. Corpus, subjects and recording procedure

The utilized corpus consists of 40 German words including nouns, adjectives, verbs and digits (Table 1), uttered by five male native German speakers aged between 28 and 36 years (average 31.8). Each subject recorded the data in eight different sessions. Between the sessions, the antennas were dismounted and remounted and the radar hardware and software were fully reset. Of the eight sessions recorded by each subject, four were recorded using the tape-based fixation method of the antennas according to [8, 17], and four were recorded using the headset proposed in this study. The goal of this experimental design was to compare the performance of the system with the headset and with tape, the latter of which is the baseline for our analysis. Care was taken to ensure that the individual antennas were positioned as consistently as possible for both conditions (headset and tape) and across sessions. The position of the antennas was about 1 cm from the corner of the mouth. Each individual session consisted of 10 repetitions of each of the 40 words, resulting in 400 recorded tokens per session. In total, the corpus consisted of 400 words × 8 sessions × 5 speakers = 16,000 tokens.

Table 1: *The word corpus used in the study and their IPA transcription [6].*

Noun		**Adjective**		**Verb**		**Digit**	
Jahr	jaːᴿ	neu	nɔœ̯	werden	vˈeːᵄdn̩	Null	nʊl
Uhr	uːᵄ	andere	ˈandəʁə	haben	hˈaːbm̩	Eins	aɛ̯ns
Prozent	pʁ̥otsˈɛnt	groß	gʁoːs	sein	zaɛ̯n	Zwei	tsv̥aɛ̯
Million	mɪli̯on	erste	ˈeːᵄstə	können	kˈœnən	Drei	dʁaɛ̯
Euro	ˈɔœ̯ʁoː	viel	fiːl	müssen	mˈʏsn̩	Vier	fiːᵄ
Zeit	tsaɛ̯t	deutsch	dɔœ̯tʃ	sollen	zɔln	Fünf	fʏnf
Tag	taːk	gut	guːt	sagen	zˈaːgŋ̍	Sechs	zɛks
Frau	fʁ̥aɔ̯	weit	vaɛ̯t	geben	gˈeːbm̩	Sieben	zˈiːbm̩
Mensch	mɛnʃ	klein	klaɛ̯n	kommen	kˈɔmən	Acht	axt
Mann	man	eigen	ˈaɛ̯g ŋ̍	wollen	vɔln	Neun	nɔœ̯n

This corpus contains exactly the same words and the same number of words per session as the corpus used for electro-optical stomatography (EOS) based silent speech recognition from [6].

A customized C++ graphical user interface was used to control the hardware, record the data and inspect them in real time. Both the radar and audio data streams were recorded simultaneously (with a fixed sample rate of 100 Hz and 44100 Hz, respectively). The detectSpeech function in MATLAB R2021a was used for word segmentation.

3.3. Classification experiments

To assess the performance of our SSI with the tape-based antenna fixation method (baseline) and with the proposed headset, we performed a classification experiment with the recorded data and used the classification accuracy as the performance metric. As classifier, we used a recurrent neural network with a simple architecture: an input layer, followed by a single bidirectional long short-term memory (BiLSTM) layer, a fully connected layer and a softmax classification layer as output.

The key hyperparameters of the used BiLSTM models are shown in Table 2. The usage of a finite validation patience value allowed "early stopping" of the training (to avoid overfitting) if the validation loss did not improve after 20 validations [17].

Table 2: *BiLSTM hyperparameters.*

Hyperparameter	Evaluated values/ranges
Number of hidden layers	1
Number of hidden units	[10, 20, 40, 60]
Max. number of epochs	200
Validation patience	20
Learning rate	0.001
Mini-batch size	8

The spectral magnitude of S1 and ΔS1 were evaluated and considered as input features for the BiLSTM network, where S1 denotes the 128 point transmission spectrum from one cheek to the other and Δ means the difference between two adjacent spectral frames. Overall, this feature set includes 256 features. In addition, other feature sets were evaluated here, including spectral magnitude of S1, spectral phase of S1 and spectral magnitude of ΔS1. All feature sets were normalized between [0-1] using the normalize function in MATLAB R2021a.

Figure 5 illustrates radar spectrograms from the cheek-to-cheek transmission path for two words spoken in two different recording sessions. It shows obvious differences between different words, and similarities between the same words uttered in different sessions.

Since the stability of the measurements is a key aspect in the development of the headset proposed in this study, we evaluated our classifier's performance with the inter-session (and intra-speaker) paradigm, in which the data used to train and test the model, came from different and non-overlapping sessions. [18]. This paradigm assesses how well the classifier recognizes patterns across sessions, i.e., after removing and reattaching of the antennas. Higher measurement stability should yield higher classification accuracies. Therefore, data from a single session were left as the hold-out test set (400 tokens) and the classifier was trained with the data from the remaining sessions (1200 tokens). The training was carried out with 80% of the training sequences (960 tokens), whereas the remaining 20% served as validation set (240 tokens). The optimization of the number of hidden units was based on the validation set's classification accuracy. Finally, the model with the highest accuracy on the validation set was tested on the test set, resulting in its classification accuracy.

For each combination of feature set, speaker and antenna fixation method we performed 20 runs of the inter-session assessment, comprising each of four different accuracies (obtained with each of the four sessions as hold out test set). Since the partition between training and validation set was random (with the MATLAB 2021a cvpartition function) and happened for each run individually, we were able to account better for the variability of this process by running several times.

4. Results

The feature set composed of the magnitude of S1 and ΔS1, achieved the highest mean accuracy among all tested feature sets and, therefore, is the one being presented here. Table 3 reports the classification results for the inter-session evaluation across all 20 repetitions for 5 speakers under the headset and tape conditions. The classification accuracies were on average an absolute 8.32% higher with the headset than with the tape i.e., 76.50% vs. 68.1%. With regard to the individual speakers, the mean accuracy with the headset was better than with tape for four of the five speakers, with an improved accuracy ranging from 2.87% (speaker 5) to 21.50% (speaker 3). Speaker 2 presented higher accuracy with tape than with headset, but by a small margin of 0.61%.

5. Discussion and Conclusion

This study investigated the inter-session word recognition accuracy of a novel radar-based SSI with a convenient and lightweight headset for a reproducible placement of the antennas on the speakers' face. While many SSI studies only investigated speech recognition performance for the intra-session paradigm [3, 15, 17], the inter-session paradigm used here is more representative of real-life applications. The results obtained in our study are comparable to those presented for a

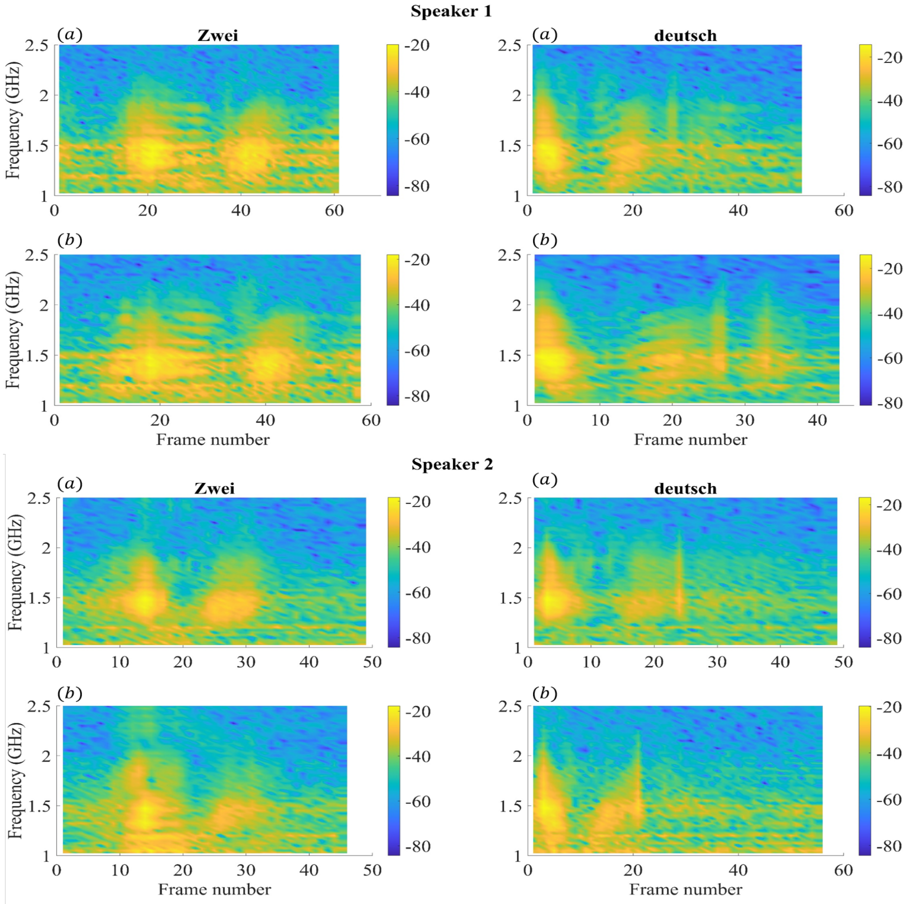

Figure 5: *Radar spectrograms for the magnitude of $\Delta S1$ (in dB) for two words ("Zwei" and "deutsch") for two speakers using the headset in two different sessions (a and b).*

Table 3: *Inter-session classification accuracy. "Mean" and "Std Dev" indicate the average and standard deviation across all test set over 20 runs, respectively.*

Speaker	**1**	**2**	**3**	**4**	**5**	**Average**
	Mean± Std Dev	Mean± Std Dev	Mean± Std Dev	Mean± Std Dev	Mean± Std Dev	
Baseline	58.95% ±5.08	69.38% ±3.00	70.05% ±3.28	64.19% ±3.25	78.35% ±6.81	68.18%
Headset	69.35% ±2.58	68.77% ±3.29	91.55% ±1.66	71.63% ±3.70	81.22% ±4.38	76.50%

sEMG-based [3] and a radar-based [17] SSI systems. The average accuracy across sessions with a smaller corpus (10 digits) and 3 speakers in [3] was 76.2%. On the other hand, [17] achieved 79.5% mean accuracy across two speakers with a narrower frequency band (1-2.5 GHz), an identical feature set, a larger corpus (50 German words) and different antenna type, while our average accuracy across five speakers was 76.50%.

Each speaker has speech idiosyncrasies and different facial features that may have played a role in our experiment. Two speakers had beards, four wore glasses and they all had different head sizes and slightly different german dialects. These factors could play a role in the high inter-speaker accuracy range (19.4% with tape and 22.78% with headset), but to what extend each of them affect the results is yet unknown.

The average accuracies obtained with the proposed headset were higher than those obtained with the taped antennas. This means that the headset has the potential to provide stable measurements, while also making the recording procedure faster and cheaper. Furthermore, the headset is portable, lightweight (only 69 g) and convenient (subjects did not report any discomfort during recording sessions).

Future work will focus on further development of the headset and designing experiments to understand which of the aforementioned idiosyncratic factors plays a role in our system's accuracy. Additional normalizing and modification of feature sets (e.g., session selection, feature space adaption, etc.) as well as recording larger corpora with more words and various speakers are also being investigated. Furthermore, the best location for the antennas has yet to be discovered, and there are other antenna types to explore for even more steady measurements.

6. Acknowledgements

The authors are grateful for financial support by the European Regional Development Fund (EFRE, project "RadarSpeech") and by the Sächsiche Aufbaubank (SAB) under the grant agreement no. 100328626. The parameter optimizations were performed at the Center for Information Services and High Performance Computing (ZIH) at TU Dresden.

7. References

[1] M. Wand and T. Schultz, “Towards real-life application of EMG-based speech recognition by using unsupervised adaptation,” in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[2] Y. Wang, T. Tang, Y. Xu, Y. Bai, L. Yin, G. Li, H. Zhang, H. Liu, and Y. Huang, “All-weather, natural silent speech recognition via machine-learning-assisted tattoo-like electronics,” *npj Flexible Electronics*, vol. 5, no. 1, 2021.

[3] L. Maier-Hein, F. Metze, T. Schultz, and A. Waibel, “Session independent non-audible speech recognition using surface electromyography,” in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2005, pp. 331–336.

[4] M. Angrick, C. Herff, E. Mugler, M. C. Tate, M. W. Slutzky, D. J. Krusienski, and T. Schultz, “Speech synthesis from ECoG using densely connected 3D convolutional neural networks,” *Journal of neural engineering*, vol. 16, no. 3, 2019.

[5] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, “Speech synthesis from neural decoding of spoken sentences,” *Nature*, vol. 568, pp. 493–498, 2019.

[6] S. Stone and P. Birkholz, “Cross-speaker silent-speech command word recognition using electro-optical stomatography,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7849–7853.

[7] F. Bocquelet, T. Hueber, L. Girin, C. Savariaux, and B. Yvert, “Real-time control of an articulatory-based speech synthesizer for brain computer interfaces,” *PLoS computational biology*, vol. 12, no. 11, 2016.

[8] P. Birkholz, S. Stone, K. Wolf, and D. Plettemeier, “Non-invasive silent phoneme recognition using microwave signals,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2404–2411, 2018.

[9] M. Pucher, N. Klingler, J. Luttenberger, and L. Spreafico, “Accuracy, recording interference, and articulatory quality of headsets for ultrasound recordings,” *Speech Communication*, vol. 123, 2020, Publisher: Elsevier, pages = 83–97,.

[10] M. J. Fagan, S. R. Ell, J. M. Gilbert, E. Sarrazin, and P. M. Chapman, “Development of a (silent) speech recognition system for patients following laryngectomy,” *Medical engineering & Physics*, vol. 30, no. 4, pp. 419–425, 2008.

[11] H. Sahni, A. Bedri, G. Reyes, P. Thukral, Z. Guo, T. Starner, and M. Ghovanloo, “The tongue and ear interface: a wearable system for silent speech recognition,” in *Proceedings of the 2014 ACM International Symposium on Wearable Computers*, 2014, pp. 47–54.

[12] L. Spreafico, “Ultrafit: A speaker-friendly headset for ultrasound recordings in speech science,” *Interspeech*, pp. 1517–1520, 2018.

[13] J. F. Holzrichter, G. C. Burnett, L. C. Ng, and W. A. Lea, “Speech articulator measurements using low power EM-wave sensors,” *The Journal of the Acoustical Society of America*, vol. 103, no. 1, 1998.

[14] A. M. Eid and J. W. Wallace, “Ultrawideband speech sensing,” *IEEE Antennas and Wireless Propagation Letters*, vol. 8, 2009.

[15] 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.

[16] P. Amini Digehsara, C. Wagner, P. Schaffer, M. Bärhold, S. Stone, D. Plettemeier, and P. Birkholz, “On the optimal set of features and the robustness of classifiers in radar-based silent phoneme recognition,” *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung*, pp. 112–119, 2021.

[17] C. Wagner, P. Schaffer, P. Amini Digehsara, M. Bärhold, D. Plettemeier, and P. Birkholz, “Silent speech command word recognition using stepped frequency continuous wave radar,” *Scientific Reports*, vol. 12, no. 1, pp. 1–12, 2022, [Online]. Available: https://doi.org/10.1038/s41598-022-07842-9.

[18] Y. Wang, M. Zhang, R. Wu, H. Gao, M. Yang, Z. Luo, and G. Li, “Silent speech decoding using spectrogram features based on neuromuscular activities,” *Brain Sciences*, vol. 10, no. 7, 2020.