ON THE OPTIMAL SET OF FEATURES AND THE ROBUSTNESS OF CLASSIFIERS IN RADAR-BASED SILENT PHONEME RECOGNITION

Pouriya Amini Digehsara[1], Christoph Wagner[1], Petr Schaffer[2], Michael Bärhold[2], Simon Stone[1], Dirk Plettemeier[2], Peter Birkholz[1]

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

[2] Chair of Radio Frequency and Photonics Engineering, Technische Universität Dresden, Germany

pouriya.amini@tu-dresden.de

Abstract: Silent speech recognition (SSR) is an active area of research with applications ranging from speech restoration to speech enhancement. Radar-based SSR has been proposed and investigated as a non-invasive method to infer vocal tract states and articulatory movements from measured changes in scattering parameters. One of the challenges in developing a radar-based SSR system is to determine the optimal set of features from these measurements. In this study, we therefore investigated the following problems: (a) The selection of the features that play the most significant role for classification. (b) The determination of the contribution of each reflection and transmission spectrum and the most important frequencies. (c) The determination of the performance of the classifiers when using fewer features. (d) The determination of the robustness of the classifiers against different noise levels. The data used in this study consisted of 230 samples of 25 German phonemes (15 vowels, each in 10 contexts, and 10 consonants, each in 8 contexts) produced by two German native speakers. Using the full feature set, a Linear Discriminant Analysis (LDA) classifier achieved up to 94 % classification accuracy for speaker 1 and 84 % for speaker 2. Using only the most important features as identified by a decision tree, the classification accuracy deteriorated slightly in most conditions, but in one case improved the accuracy from 73.5 % to 81 %. Regarding the robustness against noise, the accuracy of the LDA dropped sharply with increasing noise levels, while the decrease of the SVM's accuracy was less steep.

1 Introduction

Silent speech recognition (SSR) is an active research area to restore and enhance human speech without any acoustic signal from a speaker [1, 2]. With the help of silent speech recognition devices, it is possible to determine vocal tract states and thus recognize speech. In the medical domain, for people who have lost their larynx or otherwise cannot produce speech conventionally, SSR devices are promising communication tools. Silent speech recognition may be useful in noisy environments where automatic speech recognition does not work well [3]. SSR could also be used as a new input modality for everyday devices. Several studies have recently been carried out to develop a safe and convenient system that collects data from radar and other sensors as input sources [4-7].

For articulatory movement recognition, multiple types of sensors were used in the literature. These sensors consist of image sensors [8], electromyography (EMG) [9, 10], electromagnetic articulatory (EMA) [11, 12], permanent magnet articulography (PMA) [13] and electro-optical stomatography (EOS) [14, 15]. Most of the mentioned methods have their individual drawbacks, including invasiveness, session dependence, sensibility to sensor placement, and others. A very promising sensor technology employs non-invasive ultra-

wideband radar (UWB) sensors. Radar pulses penetrate the skin and the underlying tissue so that intra-oral vocal tract configurations can be captured from the outside [16-18].

However, only a few studies examined the potential of these radar sensors for silent speech recognition. Holzrichter et al. were the first ones to utilize a radar sensor to recognize variations of the vocal tract shape in 1998 and proposed to use these sensors in SSR devices [19]. More recently, Eid and Wallace [20] used a single radar sensor to detect 10 different digits and ascribed great potential to this method. Shin and Seo [21] studied ultra-wideband (UWB) radar sensors with two antennas to detect 10 isolated words and classified each word with an average accuracy of 85 %. Birkholz et al. [16] also presented a UWB articulation sensing system (2-12 GHz) that utilized two radar antennas placed below the chin and on the cheek of two speakers. The measured scattering parameters were used to classify different German phonemes and the results showed a 93 % and 85 % recognition rate for the two different speakers. Here, the features were the spectral coefficients of the transmission and reflection spectra with a total number of 1206. The role of each individual feature for the classification performance remained uninvestigated. It is also interesting to investigate the efficiency of the classifiers when using fewer than the 1206 features, and the robustness of the classifiers against different levels of noise.

In this study, we built on the work in [16] and used the same set of recordings. To reduce the number of features, a decision tree was used. Two established pattern recognition methods, Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM), were used to assess the classification accuracy. In the last section of this paper, the influence of different signal-to-noise ratios was examined.

2 Experimental setup and data set

The system employed two modified antipodal Vivaldi antennas attached below the chin and on the right cheek [16, 22]. During a measurement, both antennas were used to transmit and receive a frequency sweep from 2 to 12 GHz. S-parameters (scattering parameters) were used to describe how energy can propagate through a network. A network analyzer (PNA Series Network Analyzer E8364B by Agilent Technologies) recorded the reflection spectrum of each antenna and the transmission spectrum between the two antennas and calculated the scattering parameters in terms of complex spectra. As described in [16], most of the information is contained in the magnitude spectra. Therefore, only the magnitude spectra were used as features and the phase information was discarded. Each of the three-recorded spectra (one transmission, two reflection) consisted of 201 frequency points, hence the number of features was 603.

Two German native speakers were asked to record two complete sets of utterances (sustained speech sounds), both in a single session. A set consisted of 230 samples of 15 vowels (/aː, eː, iː, oː, uː, ɛː, øː, yː, ɪː, ɛ, a, ɔ, ʊ, ʏ, œ/), each in 10 different symmetric consonantal contexts, and 10 consonants (/b, d, g, l, ʁ, f, s, ʃ, m, n/), each in 8 different symmetric vowel contexts. Due to the acquisition time of the network analyzer, each target phoneme was sustained for 3 seconds. In summary, the data set of each speaker included two sets of 230 observations with 603 feature dimensions for 25 different phoneme classes.

3 Feature selection

While the results from the pilot study in [16] were already quite convincing as a proof-of-principle, the high input dimensionality is problematic for two reasons: It may limit the performance of a classifier (due to the curse of dimensionality) and it requires a long signal acquisition time, because each frequency component is measured individually. If a smaller, salient subset of these features could be identified, it would enable the use of a more specific stepped-frequency sweep to only measure at the important frequencies in much shorter time.

Therefore, we used a decision tree as a feature selection technique to identify the most important spectral components for classification.

Feature selection methods try to select the best subset of features, which have the most significant role in classification and omit the redundant attributes from a data set. In addition, it would determine the features (or some sections or frequency bands) that a classifier relies upon. This gives us a chance to identify the subset of features that is essential to separate the phoneme classes.

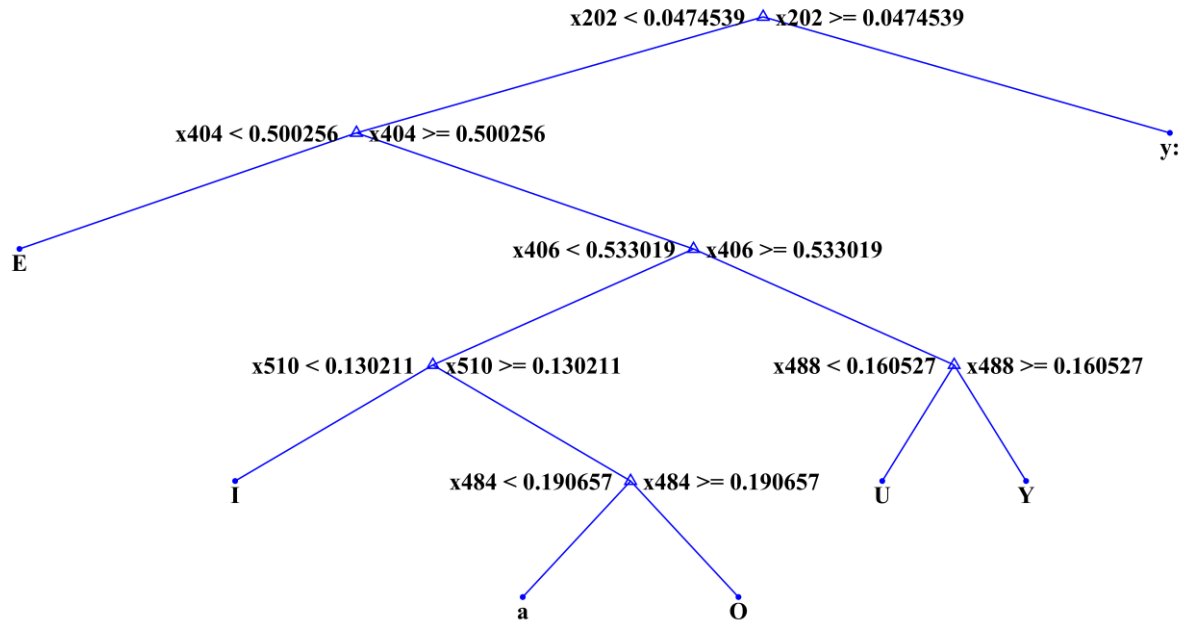

Fig.1: A part of the decision tree. Each node selects a feature to split the classes based on Gini Diversity Index (GDI). The leaf nodes correspond to the classes (phonemes).

The decision tree is one of the most well-known and simplest feature selection algorithms. It determines those features that play a role for discrimination and discards the others [23]. A decision tree tries to split classes with the minimum amount of classification error based on a Standard CART (Classification And Regression Tree) predictive model. In each node, to select the best split feature, the Gini Diversity Index (GDI) is used [24, 25]. The Gini Diversity Index is the split criterion to illustrate a node's impurity:

$$1 - \sum_{i=1}^{C} p_i^2$$

Here, C is the number of classes at the node, and p_i is the observed fraction of classes with class i that reach the node. A high purity of a node is associated with a low GDI.

This method is implemented in `MATLAB R2019b` in the `fitctree` function to fit a binary decision tree for multiclass classification. The process of creating a decision tree was repeated 30 times. For each run, a stratified holdout set of 10 % of the observations in each set was used for the final evaluation of the tuned classifier's accuracy. The remaining 90 % were used to build the decision tree and to train and tune the classifier in a cross-validation scheme. A part of the tree for speaker 1 is shown in Fig.1.

To determine the importance of each feature at each node (contribution weight), the corresponding mean square error (MSE) was calculated. At the end of the splitting process, the summation of MSE for each feature was divided by the number of branch nodes and determined the importance of each selected feature.

The average contribution weight of each selected feature in the decision trees is illustrated in Fig. 2. In Fig. 2 (a) the number of selected features for each spectrum is shown for the first repetition of speaker 1(176 features were selected). It can be seen that the most frequently selected features (92) were coefficients of the reflection spectrum of S11, and 49 features were selected from the transmission spectrum between the two antennas (S21). Also, in the reflection spectra (S11 and S22), the whole frequency band (2-12 GHz) was equally important, but for the transmission spectrum (S21) mostly the lower frequency range was important.

Fig. 2 (b) shows the cumulative sum of weights for each spectrum. For the transmission spectrum (S21), over 80% of the significant features, were in the 2-7GHz frequency band. For the reflection spectra (S11 and S22), this frequency band included only 60% and 40% of the selected features, respectively. In all three spectra, the initial part of the frequency band (2-3GHz) contained over 40% of the important features. The number of selected features were 146 for the second repetition of speaker 1. For the second speaker, the reduced feature subset contained 188 and 223 features, for the first and second repetition, respectively

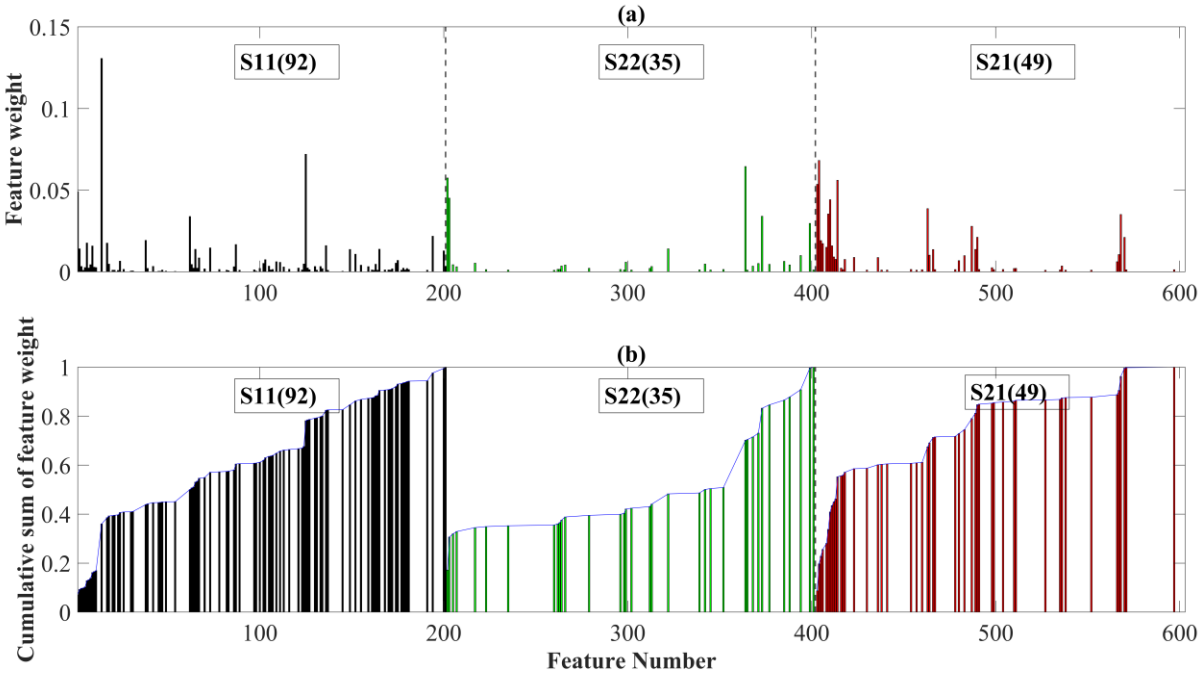

Fig.2: Feature importance; (a) the contribution weight of each feature, (b) the cumulative sum for each spectrum.

4 Classification experiments

To evaluate the impact of the feature selection on the classification accuracy, a Linear Discriminant Analysis (LDA) and a Support Vector Machine (SVM) were implemented. These methods work well with small data sets and have no (LDA) or only two (SVM with Gaussian kernel) adjustable hyperparameters. Each classifier was trained and evaluated in a speaker-dependent fashion using only a single data set from each speaker at a time.

- Linear Discriminant Analysis determines a set of basis vectors that discriminate class samples. These vectors have the minimum within-class variances and maximum between-class variances to guarantee an optimal separation of the classes. Particularly, LDA is a supervised learning model with no hyperparameter [27]. Here, we used the LDA implemented in the `MATLAB R2019b` function `fitcdiscr` using its default settings.

- A Support Vector Machine finds a hyperplane that separates two classes in an optimal sense by maximizing the margin between the hyperplane and the support vectors [28]. This is extended to multiclass classification using a one-vs-one design in the function

`fitcecoc` in `MATLAB R2019b`. For the nonlinear transformation of the features, a Gaussian kernel was used. The soft margin constant (C) and the kernel scale (σ) are the only two hyperparameters tuned in this study. These hyperparameters were determined by the internal optimization function in `fitcecoc`.

The classification accuracy for both speakers using LDA and SVM with the entire feature set and the reduced feature sets are shown in Table 1. For both recorded sets of speaker 1, all classifiers performed worse when using reduced feature sets. The best results were achieved by an SVM with a Gaussian kernel, which exhibited a loss of only 1.12% of accuracy with using only 172 of the 603 original features. For the second speaker this feature selection condition even outperformed the results on the full feature set for all classifiers. The classification accuracy of the SVM (linear kernel) when using only 32% of the original features reported the highest increase from 73.5 % to 81 %. The accuracy of LDA and the SVM (Gaussian kernel) increased by 3.04% and 6.96%, respectively.

Table.1: Classification accuracy of full and reduced feature sets with two different classifiers on the radar silent speech datasets. "Mean" and "Std Dev" indicate the average and standard deviation across all test set over 30 runs, respectively, and "NF" in parentheses means the number of the selected features over 30 runs.

Dataset	**LDA**		**SVM (linear kernel)**		**SVM (Gaussian kernel)**	
	Baseline	**Selected features**	**Baseline**	**Selected features**	**Baseline**	**Selected features**
	Mean ± Std Dev	**Mean ± Std Dev (**NF**)**	**Mean ± Std Dev**	**Mean ± Std Dev (**NF**)**	**Mean ± Std Dev**	**Mean ± Std Dev (**NF**)**
Speaker 1 repetition 1	93.84% ± 3.97	90.64% ± 5.92(176)	92.48% ± 4.95	89.36% ± 6.02(176)	92% ± 5.11	90.88% ± 5.60(176)
Speaker 1 repetition 2	90.32% ± 5.55	86% ± 6.52(146)	90.40% ± 5.05	81.04% ± 7.87(146)	89.52% ± 5.59	82.56% ± 6.85(146)
Speaker 2 repetition 1	84.32% ± 7.39	86% ± 6.52(188)	73.52% ± 8.11	81.04% ± 7.86(188)	75.6% ± 8.41	82.56% ± 6.85(188)
Speaker 2 repetition 2	75.28% ± 7.16	78.32% ± 6.71(223)	67.36% ± 8.17	69.12% ± 8.71(223)	69.12% ± 6.81	70.16% ± 8.36(223)

5 Robustness against noise

To evaluate the robustness of the tested classifiers against background noise, the test data passed through white Gaussian noise channels for signal-to-noise ratios (SNR) of 5 dB to 50 dB in 5 dB steps. For both speakers, the selected feature sets were used and the results are shown in Fig. 3. It is evident that the SVM family showed better robustness against external noise sources (with very little loss of accuracy down to around 20dB) while the LDA's accuracy dropped severely at an SNR of 35 dB or lower.

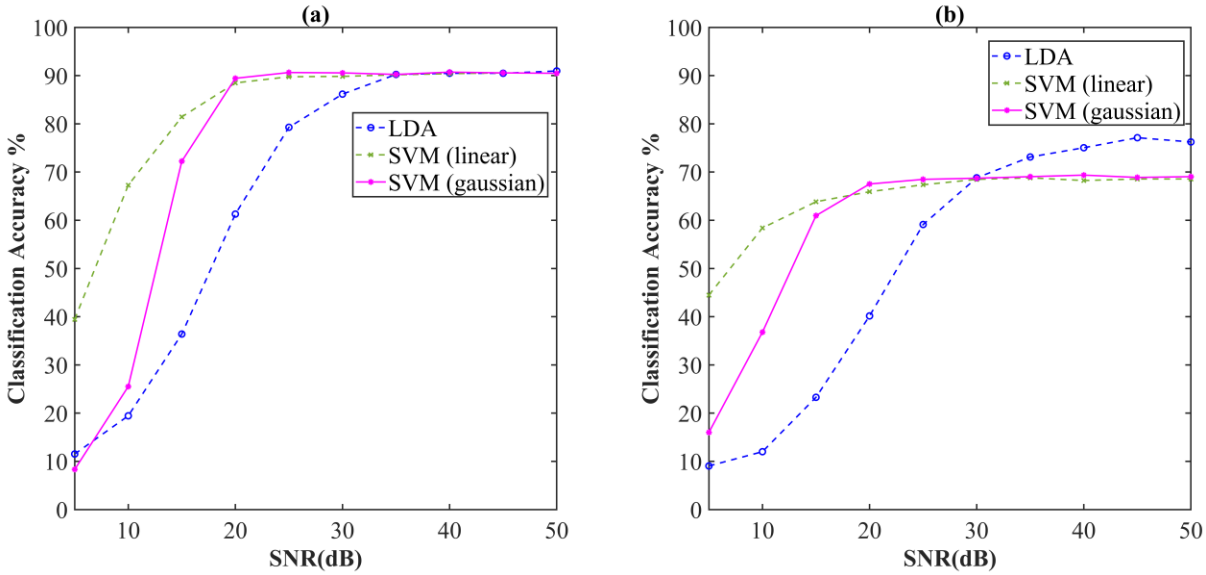

Fig. 3: Robustness of classifiers against different amount of noise; (a) classification accuracy for speaker 1 (first repetition) using the selected features (176), and (b) classification accuracy for speaker 2 (second repetition) using selected features (223).

6 Discussion and conclusion

This study tried to find the best subset of features of a high-dimensional feature vector using a decision tree and showed the importance of each feature for classification. In the previous study [16], it was investigated that the first half of the frequency band for each spectrum contained the most important features for phoneme classification. By using a decision tree, it was shown here that for the reflection spectra, the whole frequency band is important, and for the transmission spectrum over 60 % of important features are in the 2-3 GHz range. Hence, there is great potential to reduce the number of features as evidenced by the fact that the decision tree selected less than 37% of the features to build the tree. Using SVMs with different kernels, it can be expected to find even more complex relationship among features to classify the phonemes even better. By using the complete feature sets for the first speaker, SVM with linear and Gaussian kernel classified phonemes with the same accuracy as LDA, but for the second speaker, SVM scored lower than LDA. The overall accuracy of both classifiers for the second speaker dropped by over 10 %. By using selected features, the classification accuracy of all classifiers increased for the second speaker. This showed the existence of some irrelevant features in the original feature sets of the second speaker. No dental problems were reported by the first speaker, while the second speaker has a metal dental implant and four metal tooth fillings, which are likely to impede the penetration of the microwaves through the articulatory system. In the last part of the study, as noise was added to the test set, SVM with different kernels was the most noise-tolerant learning algorithm. Future work includes testing other feature selection techniques as well as dimensionality reduction techniques, e.g. evolutionary algorithms [29], and implementing a deep neural network to train with a larger amount of data from different speakers.

7 Acknowledgment

The authors gratefully acknowledge the partial financial support of this research within the project “Radar Speech” by the European Regional Development Fund (EFRE) and the Sächsiche Aufbaubank (SAB) under the grant agreement no. 100328626.

References

[1] Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J.M. and Brumberg, J.S.: *Silent speech interfaces*. In: *Speech Communication* 52, no. 4: 270-287, 2010.

[2] Hueber, T., Bailly, G. and Denby, B.: *Continuous Articulatory-to-Acoustic Mapping using Phone-based Trajectory HMM for a Silent Speech Interface*. In: *13th Annual Conference of the International Speech Communication Association* (InterSpeech 2012), Portland, United States, 2012.

[3] Lee, S. and Seo, J.: *Word Error Rate Comparison between Single and Double Radar Solutions for Silent Speech Recognition.* In: *19th International Conference on Control, Automation and Systems (ICCAS)*, pp. 1211-1214. IEEE, 2019.

[4] Rhee, J.H. and Seo, J.: *Low-cost curb detection and localization system using multiple ultrasonic sensors.* In: *Sensors* 19, no. 6, 2019.

[5] Smith, K.A., Csech, C., Murdoch, D. and Shaker, G.: *Gesture recognition using mm-wave sensor for human-car interface*. In: *IEEE Sensors Letters* 2, no. 2: 1-4, 2018.

[6] Lee, S. and Seo, J.: *IR-UWB radar-based near-field head rotation movement sensing under fixed body motions*. In: *2018 International Conference on Electronics, Information, and Communication (ICEIC)*, pp. 1-3. Ramada Plaza, Jeju, Korea, 2018.

[7] Leem, S.K., Khan, F. and Cho, S.H.: *Vital sign monitoring and mobile phone usage detection using IR-UWB radar for intended use in car crash prevention*. In: *Sensors* 17, no. 6: 1240, 2017.

[8] Ephrat, A., Halperin, T. and Peleg, S.: *Improved speech reconstruction from silent video*. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 455-462. Venice, Italy, 2017.

[9] Meltzner, G.S., Colby, G., Deng, Y. and Heaton, J.T.: *Signal acquisition and processing techniques for sEMG based silent speech recognition*. In: *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 4848-4851. Boston, Massachusetts USA, 2011.

[10] Wand, M., Schulte, C., Janke, M. and Schultz, T.: *Array-based Electromyographic Silent Speech Interface*. In: *Biosignals*, pp. 89-96, 2013.

[11] Kim, M.J., Cao, B., Mau, T. and Wang, J.: *Multiview Representation Learning via Deep CCA for Silent Speech Recognition*. In: *INTERSPEECH*, pp. 2769-2773. Stockholm, Sweden, 2017.

[12] Kim, M., Cao, B., Mau, T. and Wang, J.: *Speaker-independent silent speech recognition from flesh-point articulatory movements using an LSTM neural network*. In: *IEEE/ACM transactions on audio, speech, and language processing* 25, no. 12: 2323-2336, 2017.

[13] Hofe, R., Ell, S.R., Fagan, M.J., Gilbert, J.M., Green, P.D., Moore, R.K. and Rybchenko, S.I.: *Small-vocabulary speech recognition using a silent speech interface based on magnetic sensing*. In: *Speech Communication 55*, no. 1: 22-32, 2013.

[14] Stone, S. and Birkholz, P.: *Silent-Speech Command Word Recognition Using Electro-Optical Stomatography.* In: *INTERSPEECH*, pp. 2350-2351, 2016.

[15] Stone, S. and Birkholz, P.: *Cross-Speaker Silent-Speech Command Word Recognition Using Electro-Optical Stomatography*. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7849-7853. IEEE, 2020.

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

[17] Eid, A.M. and Wallace, J.W.: *Ultrawideband speech sensing*. In: *IEEE Antennas Wireless Propagation Letters*, Vol. 8, pp.1414-1417. Jeju, Korea, 2009.

[18] Shin, Y.H. and Seo, J.: *Towards contactless silent speech recognition based on detection of active and visible articulators using IR-UWB radar*. In: *Sensors* 16, no. 11: 1812, 2016.

[19] Holzrichter, J.F., Burnett, G.C., Ng, L.C. and Lea, W.A.: *Speech articulator measurements using low power EM-wave sensors*. In: *The Journal of the Acoustical Society of America 103*, no. 1: 622-625, 1998.
[20] Eid, A.M. and Wallace, J.W.: *Ultrawideband speech sensing*. In: *IEEE Antennas and Wireless Propagation Letters* 8: 1414-1417, 2009.
[21] Shin, Y.H. and Seo, J.: *Towards contactless silent speech recognition based on detection of active and visible articulators using IR-UWB radar*. In: *Sensors* 16, no. 11: 1812, 2016.
[22] Fang, X., Ramzan, M., Wang, Q. and Plettemeier, D.: *Compact antipodal Vivaldi antennas for body area communication*. In: *Advances in Body Area Networks I*, pp. 357-369. Springer, Cham, 2019.
[23] Safavian, S.R. and Landgrebe, D.: *A survey of decision tree classifier methodology*. In: *IEEE transactions on systems, man, and cybernetics* 21, no. 3: 660-674, 1991.
[24] Pavel, Y.P.P.A.F. and Soares, B.C.: *Decision tree-based data characterization for meta-learning*. In: *IDDM-2002*: 111, 2002.
[25] Loh, W.Y.: *Regression tress with unbiased variable selection and interaction detection*. In: *Statistica sinica* Vol. 12, No. 2 pp. 361-386, 2002.
[26] Sugumaran, V., Muralidharan, V. and Ramachandran, K.I.: *Feature selection using decision tree and classification through proximal support vector machine for fault diagnostics of roller bearing*. In: *Mechanical systems and signal processing* 21, no. 2: 930-942, 2007.
[27] Balakrishnama, S. and Ganapathiraju, A.: *Linear discriminant analysis-a brief tutorial*. In: *Institute for Signal and information Processing*, vol. 18, no, pp. 1-8, 1998.
[28] Cristianini, N. and Shawe-Taylor, J.: *An introduction to support vector machines and other kernel-based learning methods*. In: *Cambridge university press*, 2000.
[29] Digehsara, P.A., Chegini, S.N., Bagheri, A. and Roknsaraei, M.P.: *An improved particle swarm optimization based on the reinforcement of the population initialization phase by scrambled Halton sequence*. In: *Cogent Engineering*, 7(1), p.1737383, 2020.