Silent and Vocalized Speech Classification Using Surface Electromyography and Machine Learning

Millena Sartori de Oliveira[1], **Rafael Gomes Mantovani**[1] **Daniel Prado de Campos**[1]

[1]Federal University of Technology – Paraná (UTFPR), Apucarana, Paraná, Brazil

millenasartori@alunos.utfpr.edu.br,
{rafaelmantovani,danielcampos}@utfpr.edu.br

Abstract. *Communication impairments arising from medical conditions such as laryngectomies, tracheostomies, or neurological disorders significantly limit the quality of life of affected individuals. Surface electromyography (EMG) offers a promising alternative for speech recognition by capturing muscle activity without requiring audible sound production. This paper investigates the classification of silent and vocalized speech commands using EMG signals and machine learning algorithms, with the hypothesis that discriminative time-domain features are sufficient to enable effective classification even under silent speech conditions. Thirteen features were extracted from the "SilentWear" dataset, which contains recordings of eight HMI commands from four subjects, using an event-based approach across 14 EMG channels, yielding 182-dimensional feature vectors. Various dimensionality reduction techniques were applied to the feature space, and three classification algorithms were evaluated on them using Leave-One-Subject-Out (LOSO) and Leave-One-Trial-Out (LOTO) cross-validation schemes across multiple experimental scenarios. LDA achieved the best overall performance, reaching an accuracy of* 0.829 ± 0.075 *and a Macro-F1 of* 0.664 ± 0.151 *for vocalized speech under LOTO, and demonstrating competitive results even for silent speech, which highlights the discriminative power of the adopted feature set.*

1. Introduction

Communication is one of the most valuable and essential abilities developed by humanity. Throughout history, it has played a fundamental role in survival, cooperation, and social evolution. Despite its importance, many people face significant challenges in communicating due to a wide range of circumstances, including surgical procedures that affect speech production, such as tracheostomies and laryngectomies, age-related impairments that limit verbal expression, and neurological or muscular disorders that compromise the ability to speak. In such situations, conventional speech-based communication systems become ineffective, creating barriers to social interaction and reducing quality of life.

To address these challenges, researchers have explored alternative communication technologies that can interpret speech intentions without relying on audible acoustic signals. Among the available approaches, Silent Speech Recognition (SSR) systems have emerged as a promising solution, aiming to recognize words or commands directly from physiological signals associated with speech production. In particular, surface electromyography (sEMG) has attracted considerable attention because it captures the electrical activity of the facial and neck muscles involved in articulation, enabling speech

recognition even when no sound is produced. However, the development of reliable sEMG-based recognition systems remains challenging due to the noisy nature of electromyographic signals, their high variability across individuals, and the large number of features that can be extracted from them. These challenges require effective signal processing and machine learning techniques that can identify meaningful patterns and improve classification performance.

Therefore, this study investigates the classification of different spoken commands using Machine Learning (ML) algorithms, namely Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), and Random Forest (RF). The dataset employed in this work is the SilentWear dataset, introduced by [Spacone et al. 2026], which contains vocalized and silent speech recordings collected from multiple subjects using surface electromyography sensors. The main objective of this study is to evaluate the effectiveness of these classification techniques in recognizing speech commands from sEMG signals and to identify the approach that provides the best performance for silent speech recognition applications, contributing to the development of more accessible, reliable, and efficient assistive communication technologies.

2. Related Works

Silent Speech Recognition (SSR) has attracted increasing attention as an alternative communication technology for individuals who are unable to produce audible speech. Among the approaches proposed in the literature, surface electromyography (sEMG) has emerged as a promising solution for capturing the electrical activity generated by the muscles involved in speech articulation. Early studies demonstrated the feasibility of using EMG signals for speech-related applications. [Hudgins et al. 1993] proposed a feature-extraction strategy for myoelectric control that later became widely adopted in EMG-based pattern-recognition systems. Similarly, [Schultz and Wand 2010] investigated continuous speech recognition using EMG signals and highlighted the potential of muscle activity measurements for speech interfaces.

Feature extraction plays a crucial role in the performance of EMG-based recognition systems. In [Phinyomark et al. 2018], the authors presented a comprehensive review of feature extraction and selection techniques for wearable EMG sensors, emphasizing the importance of selecting discriminative features while reducing computational complexity. More recently, [Afshar et al. 2020] proposed adaptive threshold-based methods for event-driven feature extraction, demonstrating improved representation of physiological signals. Dimensionality reduction techniques are commonly employed to address the high dimensionality and redundancy present in EMG datasets. Principal Component Analysis (PCA), introduced by [Jolliffe 2002], remains one of the most widely used linear techniques for feature compression. Nonlinear approaches such as t-distributed Stochastic Neighbor Embedding (t-SNE) [van der Maaten and Hinton 2008] and Uniform Manifold Approximation and Projection (UMAP) [McInnes et al. 2018] have also been successfully applied to visualize and analyze complex high-dimensional data, often revealing structures that are not captured by linear methods.

Several ML algorithms have been explored for EMG signal classification. Support Vector Machines (SVMs), introduced by[Cortes and Vapnik 1995], are known for their effectiveness in high-dimensional classification problems, while Random Forests

(RF) [Breiman 2001] provide robust ensemble-based learning capable of handling noisy and heterogeneous data. Linear Discriminant Analysis (LDA), based on Fisher's work [Fisher 1936], remains a popular baseline classifier due to its simplicity and computational efficiency. Comparative studies have demonstrated the strengths and limitations of different classifiers when applied to EMG signals. In [Dellacasa Bellingegni et al. 2017], the authors compared LDA, SVM, multilayer perceptrons (MLP), and nonlinear logistic regression (NLR) for the classification of surface EMG signals collected from individuals with trans-radial amputation. Their results showed that although SVM achieved high classification performance, simpler approaches such as LDA remained competitive while requiring significantly lower computational complexity, highlighting the trade-off between accuracy and efficiency in EMG-based applications.

Furthermore, [Abbaspour et al. 2020] evaluated several recognition algorithms for decoding hand movements from surface EMG signals and demonstrated that classifier performance is strongly influenced by feature representation, dimensionality reduction strategies, and signal variability across subjects. Their findings emphasize the importance of systematically comparing ML algorithms when developing robust EMG-based recognition systems. These studies indicate that no single classifier consistently outperforms all others across different EMG applications, motivating further investigation of LDA, SVM, and RF classifiers for silent speech recognition tasks.

Recently, [Spacone et al. 2026] introduced "*SilentWear*", an ultra-low-power wearable platform designed for silent speech recognition using sEMG signals. The dataset provided by the authors contains both vocalized and silent speech recordings collected from multiple subjects and represents a valuable benchmark for evaluating ML algorithms in realistic wearable scenarios. Building upon this dataset, the present work investigates the performance of LDA, SVM, and RF for spoken command recognition, contributing to the understanding of classical ML algorithms in silent speech applications.

3. Experimental Methodology

Figure 1 depicts the whole experimental methodology adopted during the experiments. The next subsections give more details on them.

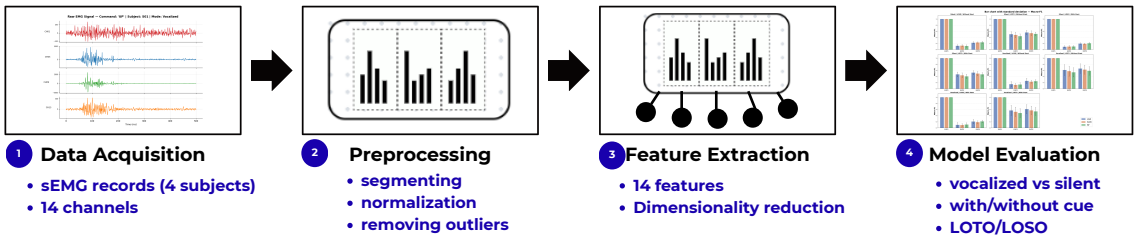


Figure 1. Silent and vocalized speech classification pipeline

3.1. Dataset

This study explored the "*SilentWear*" dataset [Spacone et al. 2026][1], a multi-session surface electromyography (sEMG) dataset designed for vocalized and silent speech recognition. This dataset comprises four participants (three males and one female), involving

[1] https://huggingface.co/datasets/PulpBio/SilentWear

eight HMI commands (up, down, left, right, start, stop, forward, and backward) in addition to a rest condition. Data acquisition was organized into three recording sessions, each comprising five vocalized and five silent batches, with 20 repetitions of each command per batch and 14 EMG channels.

3.2. Preprocessing and Feature Extraction

Feature extraction is a fundamental step in pattern recognition and ML pipelines. The input signals are transformed into a feature vector composed of descriptive attributes that capture relevant characteristics of the data. These features serve as inputs to classification algorithms, providing an informative representation of the original signals. Consequently, the performance of a classification model is strongly influenced by the quality and discriminative power of the extracted features [Bishop 2006].

Signal processing can be performed using a variety of techniques. One of the most widely used approaches is the windowing technique: the raw signal is divided into fixed-length temporal segments prior to feature extraction. However, there is no confirmation that the pre-established window boundaries will coincide with the natural boundaries of speech events. In light of this, relevant information may be distributed across different segments, which could also affect the quality of the extracted features and the classification model's performance. For this reason, instead of using windowing, the recordings were segmented into events according to changes in the provided class labels. Consecutive samples sharing the same *Label str* were considered part of the same event. Thus, each event corresponds to a continuous labeled segment of the recording, and feature extraction was performed over the complete event.

For this experiment, thirteen (13) features were extracted from each EMG channel. This feature set includes: the four classical time-domain features introduced by [Hudgins et al. 1993], the TD9 feature set proposed by [Phinyomark et al. 2018], and also the standard deviation (STD) as an additional measure of signal dispersion. These features were chosen to provide a broad representation of the EMG signal, including information on amplitude, variability, complexity, and temporal characteristics. All these features are listed in Table 1.

To avoid including transition segments and potentially unreliable segments, events shorter than 200 samples (equivalent to 400ms) were discarded. Before feature extraction, each EMG channel was normalized independently within each event using z-score normalization. For each channel, the mean and standard deviation were calculated from the samples belonging to that event, and the signal was transformed to have zero mean and unit variance.

Because each EMG channel was independently normalized within each event using z-score normalization, the absolute scale and dispersion characteristics of the original signal are substantially affected before feature extraction. Consequently, features such as RMS, STD and VAR may provide limited additional discriminative information. Therefore, an ablation analysis was performed to assess the contribution of these features to classification performance. Two feature representations were considered: the complete feature set, containing all 13 features per channel (182 features), and an ablated representation excluding RMS, STD and VAR, resulting in 10 features per channel (140 features).

Table 1. Extracted EMG features and their descriptions.

N.	Feature	Description	Information Captured
1	RMS	Root Mean Square	Signal power and muscle activation intensity
2	MAV	Mean Absolute Value	Average signal amplitude
3	STD	Standard Deviation	Signal variability and dispersion
4	WL	Waveform Length	Signal complexity and waveform changes
5	ZC	Zero Crossings	Frequency-related information
6	SSC	Slope Sign Changes	Frequency-related information and waveform variations
7	IAV	Integrated Absolute Value	Total muscle activity over time
8	VAR	Variance	Signal power and variability
9	DASDV	Difference Absolute Standard Deviation Value	Variability of consecutive signal differences
10	MSR	Mean Square Root	Signal amplitude characteristics
11	MFL	Maximum Fractal Length	Signal complexity and nonlinear behavior
12	WAMP	Willison Amplitude	Frequency and muscle contraction activity
13	LS	L-Score	Distribution and amplitude characteristics

3.3. Dimensionality Reduction

Before the classification task, dimensionality reduction (DR) techniques were used to visualize the feature space and assess class separability. Since each event was represented by a feature vector containing 182 dimensions, three DR techniques were used to guide visualization analysis: i) Principal Component Analysis (PCA); ii) t-distributed Stochastic Neighbor Embedding (t-SNE); iii) and Uniform Manifold Approximation and Projection (UMAP). PCA projects the original feature space onto a smaller set of components while retaining as much variance as possible [Jolliffe 2002]. On the other hand, t-SNE accentuates local relationships among samples, which helps visualize potential clusters [van der Maaten and Hinton 2008]. UMAP maintains this idea: it preserves both local and global structures of the data [McInnes et al. 2018].

PCA, t-SNE, and UMAP were applied to project the data into two dimensions for visualization and qualitative assessment of class separability. *These DR techniques were used exclusively for visualization and were not included in the classification pipeline*. To reduce computational cost and maintain class balance, a stratified subsampling procedure was performed before t-SNE and UMAP. It is worth noting that PCA was computed using the complete feature set, whereas the stratified subsample was used exclusively for the subsequent t-SNE and UMAP projections, as well as for the corresponding PCA visualization, in order to keep the three projections directly comparable in the resulting plots.

3.4. ML Algorithms

In this study, three classification algorithms were evaluated: Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), and Random Forest (RF). LDA seeks projections that maximize class separability and minimize within-class variance, while SVM defines decision boundaries by maximizing the margin between classes [Fisher 1936, Cortes and Vapnik 1995]. Lastly, RF combines multiple decision trees (DTs) to improve robustness and generalization [Breiman 2001]. These methods were selected because they have been successfully applied to EMG pattern recognition tasks [Phinyomark et al. 2018]. In specific, LDA was implemented with the default singular value decomposition (SVD) solver; SVM used a radial basis function (RBF) kernel with $C = 1$ and γ set to 'scale'; and RF was configured with 100 estimators. All values

correspond to the default settings of the `scikit-learn` library, and no hyperparameter optimization was performed.

3.5. Learning Scenarios

Numerous dataset configurations and classification scenarios were considered to analyze the performance of the chosen classification algorithms. During data acquisition, participants had to pronounce the command *start* before each target word: this command is known as a protocol cue, a prearranged signal that indicates the beginning of a task and also ensures consistency. Since repeating the cue may introduce patterns unrelated to the target commands, two data set configurations were considered: one that includes the protocol cue and another that excludes it. When included, the protocol cue was treated as an additional target class within the classification task, rather than as noise embedded in the neighboring command segments. Consequently, the "*With Start*" configuration involves one extra class compared to the "*Without Start*" configuration for the same experimental scenario.

For each configuration, three classification scenarios were considered. For the first one (EXP1), a binary classification problem involving speech and rest was expected - the main objective is to examine the ability of the classification process in distinguishing between active speech and rest periods. The second (EXP2) consisted of only speech commands, so the rest of the class was excluded from the configuration. As a result, the three algorithms were tasked with multi-class classification. For the third (EXP3) scenario, all speech commands and the rest condition were combined, representing the most challenging scenario.

Each scenario was subdivided into vocalized and silent speech. Furthermore, for each configuration, there were two validation protocols:

- **Leave-One-Subject-Out (LOSO)**: data from three subjects were used for training, while the remaining subject was reserved for testing;
- **Leave-One-Trial-Out (LOTO)**: the evaluation was performed separately for each subject, with one trial used as the test set and all remaining trials from the same subject used for training.

For the LOTO protocol, folds in which the test trial contained a class absent from the corresponding training trials were discarded, since such cases cannot be evaluated fairly by the classifier. While LOTO focuses on measuring the consistency of the classifiers on the same subjects, LOSO is a more challenging evaluation because it assesses the algorithm's ability to generalize to unseen subjects. Table 2 shows the details of the different learning tasks explored in the experiments.

To avoid data leakage, outlier removal was performed independently within each cross-validation fold. For each training fold, the maximum absolute feature value was computed for each event, and the 99th percentile of these values was used as the outlier threshold. Events exceeding this threshold were removed from the training set. The same threshold estimated exclusively from the training data was then applied to the corresponding test set. After outlier removal, feature standardization was performed separately within each fold. The *StandardScaler* was fitted exclusively on the training data and the resulting transformation was applied to the corresponding test data.

Table 2. Detailed comparison of experimental configurations (EXP1, EXP2, and EXP3).

Configuration Aspect	EXP1	EXP2	EXP3	Note
		Dataset Structure		
Number of Classes	2	8	9	Progressive
Classes Definition	Speech, REST	8 Commands	Commands + REST	Increasing
		Segmentation		
Segments/Class	2,400	2,400	2,400	Balanced
Segments/Subject	600	600	600	Equal
Total Segments	4,800	19,200	21,600	4× and 4.5× EXP1
		Experimental Properties		
Classification	Binary	8-class	9-class	Complexity
Task Difficulty	Low	Moderate	High	Increasing
Separation	Easy	Challenging	Very Hard	Command discrimination
REST Handling	Included	Excluded	Included	Strategy diff
		Evaluation		
Class Balance	Balanced	Balanced	Balanced	All equal
Validation	LOSO/LOTO	LOSO/LOTO	LOSO/LOTO	Same protocols
Expected Performance	High	Moderate	Lower	Difficulty impact

The chosen metrics for interpreting the classifier's performance were accuracy and Macro-F1 score: the first measures the overall proportion of correctly classified samples, whereas the latter accounts for the balance between precision and recall across all classes. All three classifiers were implemented with the Scikit-learn library. To establish a fair comparison between classification methods, only the defaults hyperparameter values were used, and no optimization was performed. Finally, all the codes and datasets used in this study are publicly available[2].

4. Results and Discussions

4.1. 2-Dimensional Reduced Projections

When analyzing DR results, it is common to examine the presence of clusters: groups of samples that share similar characteristics and are close to one another in the feature space [Bishop 2006]. The emergence of distinct clusters may provide qualitative evidence of the complexity of separating classes and help evaluate whether the feature set captured discriminative information.

Figure 2 shows the 2D data visualization obtained by the DR techniques. For every method applied, it is possible to clearly distinguish clusters corresponding to the rest class (gray points) and several other colored clusters associated with spoken commands. The visual separation between these two sample categories might suggest that, given the implemented feature set and the complexity of the problem, the classification models later applied may be able to discriminate between rest and speech with relative ease.

[2]https://github.com/ic2d/silent_and_vocalized_speech_recognition

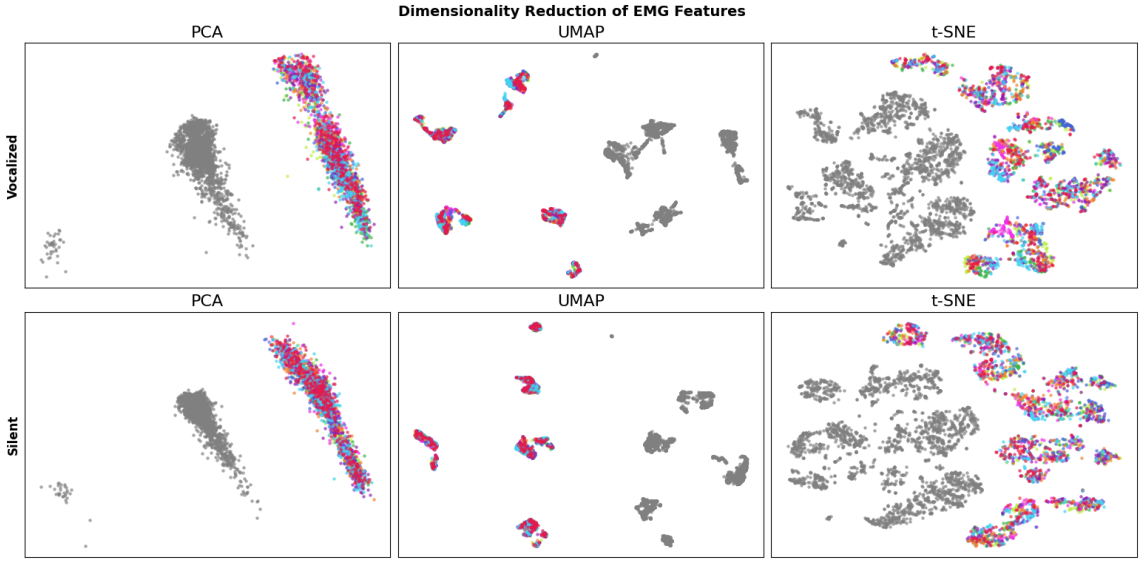


Figure 2. Dimensionality reduction projections of the extracted feature vectors

In a more detailed analysis, it is important to note that clusters that embrace spoken commands rarely consist of a single class. Although the data set is composed of eight (8) different directions, words may share similar acoustic and phonetic characteristics: overlapping phoneme sequences, comparable articulatory patterns, or even similar spectral features. As a result, their representations can occupy nearby regions in the feature space [Rabiner and Juang 1993]. Thus, the presence of clusters containing several classes is already expected and might also suggest that some commands are more difficult to separate based solely on the extracted speech features.

A clear difference can be observed between the DR techniques. PCA separates speech from the rest of the samples, but it fails to provide more detailed discrimination among speech commands, given that they remain highly overlapping. By contrast, UMAP and t-SNE provide a clearer visualization of the results: within the speech data, several sub-clusters reveal structures that could not be observed in the linear projection. Among the nonlinear projections, t-SNE provided the most visually distinct sub-clusters, whereas UMAP also revealed structure among the speech commands. The observed structure might suggest that linear projections are insufficient to fully capture the relationships between the extracted features and class labels.

4.2. Classification Results

The first experiment (EXP1) achieved near-perfect performance across all classifiers, speech modes, Start configurations, and validation protocols - revealing that the discrimination between speech and rest periods may represent a somewhat simple classification task for the chosen algorithms. Previous DR analyses revealed a coherent separation between rest and speech samples in the feature space, which suggested the possibility of a clear distinction. Therefore, the classification results support the cluster structures observed in the DR projections. To avoid redundancy, the complete results table for this experiment has been omitted.

Tables 3 and 4 present the complete results for the other two experiments. For the second (EXP2) and third experiments (EXP3), a clear performance difference was

Table 3. EXP2 – Classification between speech commands, comparing the ablated (10 features) and complete (13 features) feature sets. Values are reported as mean ± standard deviation.

				10 Features		13 Features (All)	
Classifier	**Mode**	**Start Signal**	**CV**	**Accuracy**	**Macro-F1**	**Accuracy**	**Macro-F1**
LDA	Silent	With Start	LOSO	0.150 ± 0.030	0.092 ± 0.035	0.151 ± 0.030	0.093 ± 0.034
			LOTO	0.488 ± 0.076	0.480 ± 0.078	0.480 ± 0.075	0.471 ± 0.077
		Without Start	LOSO	0.177 ± 0.035	0.124 ± 0.038	0.177 ± 0.035	0.123 ± 0.037
			LOTO	0.516 ± 0.081	0.508 ± 0.084	0.512 ± 0.081	0.503 ± 0.085
	Vocalized	With Start	LOSO	0.169 ± 0.075	0.100 ± 0.077	0.167 ± 0.073	0.098 ± 0.077
			LOTO	0.581 ± 0.162	0.573 ± 0.169	0.579 ± 0.160	0.571 ± 0.167
		Without Start	LOSO	0.216 ± 0.081	0.151 ± 0.087	0.214 ± 0.082	0.150 ± 0.089
			LOTO	$\mathbf{0.626} \pm 0.170$	$\mathbf{0.618} \pm 0.177$	0.622 ± 0.169	0.614 ± 0.176
RF	Silent	With Start	LOSO	0.151 ± 0.010	0.107 ± 0.029	0.149 ± 0.012	0.102 ± 0.027
			LOTO	0.417 ± 0.068	0.401 ± 0.069	0.421 ± 0.068	0.405 ± 0.068
		Without Start	LOSO	0.166 ± 0.013	0.114 ± 0.033	0.168 ± 0.020	0.118 ± 0.037
			LOTO	0.454 ± 0.070	0.439 ± 0.070	0.455 ± 0.074	0.438 ± 0.075
	Vocalized	With Start	LOSO	0.166 ± 0.055	0.116 ± 0.055	0.162 ± 0.048	0.109 ± 0.055
			LOTO	0.502 ± 0.142	0.487 ± 0.147	0.501 ± 0.142	0.485 ± 0.148
		Without Start	LOSO	0.194 ± 0.052	0.132 ± 0.056	0.198 ± 0.063	0.138 ± 0.065
			LOTO	0.545 ± 0.156	0.531 ± 0.160	0.553 ± 0.146	0.538 ± 0.151
SVM	Silent	With Start	LOSO	0.153 ± 0.029	0.101 ± 0.043	0.151 ± 0.031	0.100 ± 0.039
			LOTO	0.470 ± 0.076	0.458 ± 0.078	0.453 ± 0.076	0.439 ± 0.079
		Without Start	LOSO	0.182 ± 0.024	0.129 ± 0.036	0.181 ± 0.032	0.130 ± 0.034
			LOTO	0.507 ± 0.083	0.496 ± 0.087	0.492 ± 0.085	0.480 ± 0.089
	Vocalized	With Start	LOSO	0.154 ± 0.043	0.085 ± 0.045	0.158 ± 0.049	0.092 ± 0.048
			LOTO	0.545 ± 0.145	0.533 ± 0.151	0.536 ± 0.147	0.525 ± 0.153
		Without Start	LOSO	0.187 ± 0.044	0.118 ± 0.041	0.192 ± 0.051	0.125 ± 0.048
			LOTO	0.592 ± 0.156	0.582 ± 0.163	0.585 ± 0.155	0.575 ± 0.161

observed between the vocalized and silent conditions. Across all evaluated classifiers, vocalized speech achieved higher accuracy and Macro-F1 scores than silent speech. Those results might suggest that vocalized speech can provide more discriminative information than silent speech, which facilitates the separation between different commands. Also, the LOTO protocol achieved higher performance than LOSO. This behavior is consistent with previous findings in EMG-based speech recognition: as reported by[Schultz and Wand 2010], one of the major challenges in EMG speech recognition is speaker dependency, since elements such as speaking style and rate, pronunciation, and physiological differences might affect the recorded signals. In addition, variations in electrode positioning and tissue properties might also increase signal variability. Consequently, better performance may be achieved when training and testing data are derived from the same subject. The substantial performance gap between LOTO and LOSO is consistent with the presence of subject-dependent characteristics in the learned representations. As a result, although the models are capable of learning discriminative patterns, it remains difficult to generalize to unseen subjects.

Despite including an additional class, EXP3 consistently achieved higher accuracy and Macro-F1 than EXP2: for the LDA classifier under the vocalized, without-start condition and LOSO cross-validation, EXP3 achieved an accuracy of 0.829 ± 0.075 and a Macro-F1 score of 0.664 ± 0.151. Under the same conditions, EXP2 yielded an accuracy

Table 4. EXP3 – Classification considering all classes, comparing the ablated (10 features) and complete (13 features) feature sets. Values are reported as mean ± standard deviation.

				10 Features		13 Features (All)	
Classifier	**Mode**	**Start Signal**	**CV**	**Accuracy**	**Macro-F1**	**Accuracy**	**Macro-F1**
LDA	Silent	With Start	LOSO	0.584 ± 0.020	0.198 ± 0.024	0.585 ± 0.017	0.196 ± 0.025
			LOTO	0.745 ± 0.037	0.530 ± 0.068	0.745 ± 0.037	0.526 ± 0.068
		Without Start	LOSO	0.625 ± 0.016	0.226 ± 0.032	0.624 ± 0.017	0.223 ± 0.036
			LOTO	0.776 ± 0.037	0.563 ± 0.073	0.775 ± 0.037	0.560 ± 0.073
	Vocalized	With Start	LOSO	0.590 ± 0.034	0.206 ± 0.068	0.591 ± 0.035	0.208 ± 0.070
			LOTO	0.794 ± 0.077	0.619 ± 0.147	0.791 ± 0.078	0.614 ± 0.149
		Without Start	LOSO	0.640 ± 0.035	0.257 ± 0.084	0.639 ± 0.037	0.257 ± 0.086
			LOTO	$\mathbf{0.829} \pm 0.075$	$\mathbf{0.664} \pm 0.151$	0.827 ± 0.076	0.660 ± 0.155
RF	Silent	With Start	LOSO	0.585 ± 0.011	0.211 ± 0.021	0.584 ± 0.015	0.208 ± 0.031
			LOTO	0.711 ± 0.032	0.464 ± 0.058	0.716 ± 0.033	0.470 ± 0.058
		Without Start	LOSO	0.624 ± 0.011	0.236 ± 0.030	0.626 ± 0.010	0.239 ± 0.037
			LOTO	0.753 ± 0.033	0.513 ± 0.062	0.754 ± 0.032	0.514 ± 0.059
	Vocalized	With Start	LOSO	0.587 ± 0.016	0.211 ± 0.033	0.590 ± 0.023	0.212 ± 0.044
			LOTO	0.758 ± 0.069	0.548 ± 0.131	0.753 ± 0.067	0.539 ± 0.130
		Without Start	LOSO	0.632 ± 0.028	0.244 ± 0.054	0.632 ± 0.029	0.247 ± 0.053
			LOTO	0.796 ± 0.066	0.597 ± 0.131	0.792 ± 0.069	0.588 ± 0.140
SVM	Silent	With Start	LOSO	0.581 ± 0.011	0.191 ± 0.023	0.582 ± 0.008	0.190 ± 0.022
			LOTO	0.730 ± 0.038	0.499 ± 0.071	0.731 ± 0.039	0.495 ± 0.071
		Without Start	LOSO	0.622 ± 0.009	0.231 ± 0.017	0.623 ± 0.007	0.233 ± 0.019
			LOTO	0.766 ± 0.038	0.540 ± 0.075	0.766 ± 0.039	0.537 ± 0.077
	Vocalized	With Start	LOSO	0.584 ± 0.024	0.186 ± 0.042	0.586 ± 0.026	0.187 ± 0.045
			LOTO	0.772 ± 0.071	0.575 ± 0.137	0.770 ± 0.071	0.573 ± 0.139
		Without Start	LOSO	0.630 ± 0.026	0.226 ± 0.044	0.631 ± 0.026	0.230 ± 0.048
			LOTO	0.811 ± 0.071	0.628 ± 0.144	0.809 ± 0.071	0.625 ± 0.145

of 0.626 ± 0.170 and a Macro-F1 score of 0.618 ± 0.177, respectively. This apparent improvement is explained by the strong separability of the rest class observed in the DR analysis. Thus, the inclusion of an easily distinguishable class increases the overall classification performance, even though discrimination among individual speech commands remains challenging.

Accuracy and Macro-F1 scores were consistently lower when the Start signal was included (*With Start*). This results suggests that the presence of the protocol cue may introduce additional variability or information that is not directly related to target-command discrimination. It should be noted, however, that the this configuration also involves one additional class to be discriminated, which may partly account for the observed performance decrease independently of any noise introduced by the cue itself.

It is also worth highlighting the surprising and competitive performance of LDA algorithm: despite being considered a relatively simple classifier, it achieved comparable, and in several scenarios superior, results compared to SVM and RF. This finding is consistent with the work of [Dellacasa Bellingegni et al. 2017], who compared different learning algorithms on sEMG data and reported that LDA achieved competitive classification performance while maintaining significantly lower computational complexity. The authors argued that when the extracted features provide sufficient discriminative informa-

tion, simpler classifiers can perform as well as more sophisticated models. The results obtained in the present study support this observation, particularly under the LOTO protocol, where LDA consistently achieved the highest accuracy and Macro-F1 scores.

Furthermore, [Abbaspour et al. 2020] demonstrated that classifier performance in sEMG-based recognition tasks strongly depends on feature representation and signal variability. According to their findings, improvements in feature extraction and DR may have a greater impact on recognition accuracy than increasing classifier complexity. This observation aligns with the results reported here, where the relatively simple LDA classifier frequently matched or outperformed more complex approaches. Therefore, the results suggest that the proposed feature extraction strategy successfully captured relevant information from the EMG signals, thereby creating a feature space in which class discrimination could be achieved effectively without relying on highly nonlinear decision models.

Although accuracy reached relatively high values in some scenarios, the corresponding Macro-F1 scores were considerably lower, for example, 0.826 accuracy against 0.660 Macro-F1 under the LOTO protocol, without the Start signal, for vocalized speech using LDA. This discrepancy is particularly pronounced in EXP3 and is explained by the inclusion of the rest class: since rest samples are easily distinguishable from speech commands, classifiers achieve high recognition of that class, leading to a relatively high overall accuracy, while recognition of individual speech commands remains modest. The Macro-F1 score, by averaging performance equally across all classes regardless of their frequency, provides a more honest evaluation of the classification difficulty. This behavior suggests that classification performance is not uniformly distributed across all classes and that some commands share similar feature representations, making their discrimination more challenging, consistent with the overlapping clusters observed in the t-SNE projections.

5. Conclusions

This paper investigated the classification of silent and vocalized speech using EMG signals and different classification algorithms. DR analyses showed a clear separation between rest and speech samples, which was consistent with the near-perfect classification performance observed in the speech-versus-rest experiment. The visualizations also highlighted substantial overlap among command classes, indicating that distinguishing individual commands is considerably more challenging than separating speech from rest.

The performance of the three chosen classification algorithms (LDA, SVM, and RF) was evaluated across three experimental scenarios. Overall, the induced models showed comparable performance, with LDA achieving the highest accuracy and Macro-F1 values in several experimental conditions, particularly under the LOTO protocol: Macro-F1 of 0.616 in EXP2 and 0.664 in EXP3. The results also demonstrated that vocalized speech generally achieved higher classification performance than silent speech. Additionally, removing the protocol cue signal generally improved classification performance across the evaluated scenarios. Furthermore, LOSO was consistently more challenging than LOTO, suggesting that the extracted features capture not only command-related information but also subject-specific characteristics. Future work should focus on improving subject-independent performance and investigating feature extraction strategies that reduce inter-subject variability while preserving discriminative information.

Use of artificial intelligence: During the writing of this document, we used the *Claude Opus 4.8 LLM* for grammatical correction and text organization. We also take full responsibility for all content, including checking for plagiarism and correcting all text.

References

Abbaspour, S., Lindén, M., Gholamhosseini, H., Naber, A., and Ortiz-Catalan, M. (2020). Evaluation of surface emg-based recognition algorithms for decoding hand movements. *Medical & Biological Engineering & Computing*, 58(1):83–100.

Afshar, S., Ralph, N., Xu, Y., Tapson, J., van Schaik, A., and Cohen, G. (2020). Event-based feature extraction using adaptive selection thresholds. *Sensors*, 20(6):1600.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Dellacasa Bellingegni, A., Gruppioni, E., Colazzo, G., Davalli, A., Sacchetti, R., Guglielmelli, E., and Zollo, L. (2017). Nlr, mlp, svm, and lda: A comparative analysis on emg data from people with trans-radial amputation. *Journal of NeuroEngineering and Rehabilitation*, 14(1):82.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188.

Hudgins, B., Parker, P., and Scott, R. N. (1993). A new strategy for multifunction myoelectric control. *IEEE Transactions on Biomedical Engineering*, 40(1):82–94.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer Series in Statistics. Springer, New York, 2 edition.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Phinyomark, A., Khushaba, R. N., Ibáñez-Marcelo, E., Patania, A., Scheme, E., and Petri, G. (2018). Feature extraction and selection for myoelectric control based on wearable emg sensors. *Sensors*, 18(5):1615.

Rabiner, L. R. and Juang, B.-H. (1993). *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, New Jersey.

Schultz, T. and Wand, M. (2010). Modeling coarticulation in emg-based continuous speech recognition. *Speech Communication*, 52(4):341–353.

Spacone, G., Frey, S., Pollo, G., Burrello, A., Pagliari, D. J., Kartsch, V., Cossettini, A., and Benini, L. (2026). Silentwear: an ultra-low power wearable system for emg-based silent speech recognition. *arXiv preprint arXiv:2603.02847*.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.