

# Whistler Identification in Whistled Spanish (Silbo): A Case Study

Alejandro López-García<sup>1</sup>[0009–0009–9161–2590], María  
Alfaro-Contreras<sup>1</sup>[0000–0003–1676–3101], Julien Meyer<sup>2</sup>[0000–0002–4747–1884], and  
Jose J. Valero-Mas<sup>1</sup>[0000–0001–8667–4070]

- <sup>1</sup> Pattern Recognition and Artificial Intelligence Group, University of Alicante, Spain  
alg166@gcloud.ua.es, {malfaro, jjvalero}@dlsi.ua.es  
<sup>2</sup> Université Grenoble Alpes, CNRS, GIPSA-Lab, Grenoble, France  
julien.meyer@cnrs.fr

**Abstract.** Deemed one of the world’s most representative whistled languages, the Canary Islands’ whistled Spanish, locally known as Silbo, has long attracted linguistic research. However, most studies have adopted linguistic, ethnological, or bioacoustic perspectives, overlooking the potential of computational methods within the digital humanities. This work advances the computational study of Silbo by presenting the first automated approach to Speaker Identification (SI)—*i.e.*, the process of determining the speaker of a given utterance by computational means—in a closed-set configuration for this language. The proposal leverages standard feature extraction methods as well as pre-trained Speech Recognition models to extract representative embeddings and incorporates class-balancing mechanisms to mitigate biases arising from uneven representation of whistlers in the data—*i.e.*, label imbalance. The results obtained on the only existing dataset specifically designed for computational analysis of Silbo, comparing three representative feature extraction methods, three oversampling policies, and five classification strategies, validate the proposal, achieving  $F_1$  scores close to 90% in the best-case scenarios. While laying a solid foundation for SI in Silbo, this study also highlights the scarcity of computational research on whistled languages, and particularly Silbo, emphasizing the need for further work to bridge traditional linguistic research and modern digital humanities.

**Keywords:** Whistled languages, Speaker identification, Silbo, Speech processing.

## 1 Introduction

Whistled languages serve as a unique means of communication, in which regular spoken speech is replaced by modulated whistles that retain the same linguistic content while still enabling high intelligibility levels [20]. These languages are particularly useful in environments where spoken speech is ineffective due to challenging orographic conditions, such as long distances or dense jungles [3].

Nowadays, approximately 80 languages are known to have developed this speech type, but far fewer are regularly used worldwide [19].

The whistled Spanish of the Canary Islands, locally known as Silbo, stands as the most widely used whistled language in the world [32]. This prominence owes much to sustained promotional efforts by the regional government and cultural organizations, such as the “Asociación Cultural y de Investigación de lenguajes silbados Yo Silbo”<sup>3</sup>, and to the inclusion of one of its variants (Silbo of La Gomera) on UNESCO’s Representative List of the Intangible Cultural Heritage of Humanity in 2009.<sup>4</sup>

Due to its cultural significance, Silbo has been a subject of research, including geolinguistic analyses of whistled speech [21], works on bioacoustics comparing whistles with communication among animals [22], and a large panel of psycholinguistic investigations, including studies exploring the relationship between musical knowledge and language proficiency [27] or comprehension analyses among expert practitioners of the tradition [23]. Nevertheless, these works have thereby neglected other aspects of the speech field, such as the computational one. It must be remarked that, while the related literature comprises seminal works related to low-level procedures for extracting signal-based descriptors from whistled speech [14], the study by Jakubiak [13] on the automated transcription of Silbo constitutes the first proposal focusing on the high-level analysis of this language by computational means. More recently, O’Brien and Marczyk [28] considered the data assortment presented in this latter work to differentiate modal and whistled speech with computational approaches.

In this work, we further contribute to the study of whistled Spanish from a computational perspective. More precisely, we present the first approach to Speaker Identification (SI) applied to Silbo, *i.e.*, the process of identifying the speaker—in our case, the whistler—of a given utterance through computational means [12]. Notably, this task proves to be remarkably relevant for the computational analysis of speech signals since, beyond biometric purposes [25], SI may enhance the accuracy of multi-speaker transcription systems by adapting the recognition framework to the characteristics of each individual speaker [2].

Our SI proposal for whistled Spanish leverages standard feature extraction methods as well as pre-trained Speech Recognition models to extract representative embeddings from the utterances, which are then post-processed and adapted for their eventual identification using classification systems. The results obtained using different state-of-the-art transcription models, data-balancing methods, and classification strategies on the Jakubiak dataset [13] prove the validity of our approach, achieving remarkably competitive classification performance in particular configurations. These findings support the effectiveness of the proposed method and lay the groundwork for future research in this field, contributing to the broader effort of preserving and raising awareness of these unique forms of communication.

<sup>3</sup> <http://www.yosilbo.com>

<sup>4</sup> <https://ich.unesco.org/en/RL/whistled-language-of-the-island-of-la-gomera-canary-islands-the-silbo-gomero-00172>

The remainder of this manuscript is structured as follows: Section 2 introduces the recognition framework developed for this task; Section 3 describes the experimental setup; Section 4 presents and discusses the obtained results; and finally, Section 5 concludes the work and outlines potential future research directions.

## 2 Methodology

This section formalizes the Speaker Identification (SI) problem and presents the methodology proposed to address this task. Note that, in this work, the SI problem is modeled as a multiclass classification task in which a query utterance must be identified as produced by one of the whistlers from a fixed set of candidates, *i.e.*, a closed-set identification framework.

Let  $\mathcal{X}$  and  $\mathcal{C}$  respectively denote the spaces of Silbo recordings and their associated labels (*i.e.*, the actual whistlers of the utterances), related by the function  $\Omega : \mathcal{X} \rightarrow \mathcal{C}$ . The goal of the SI task is to approximate this function as accurately as possible by learning an estimate  $\hat{\Omega}$  using a set of labeled data,  $\mathcal{T} \subset \mathcal{X} \times \mathcal{C}$ . To achieve this, we propose the scheme illustrated in Fig. 1, which is described below.

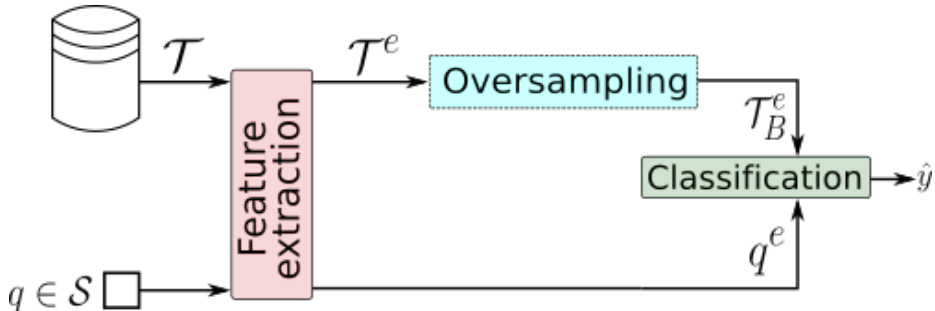


Fig. 1: Graphical description of the scheme proposed for the Speaker Identification task for Silbo speech.

The labeled Silbo recordings,  $\mathcal{T} = \{(x_i, y_i) : x_i \in \mathcal{X}, y_i \in \mathcal{C}\}_{i=1}^{|\mathcal{T}|}$ , initially undergo a feature extraction or embedding stage—denoted as the *Feature extraction* phase in the figure—resulting in the set  $\mathcal{T}^e = \{(x_i^e, y_i) : x_i^e \in \mathbb{R}^f\}_{i=1}^{|\mathcal{T}^e|}$  of embedded data, where  $f$  denotes the size of the feature representation. To mitigate potential biases, the embedded representations  $\mathcal{T}^e$  are then artificially balanced in the *Oversampling* process, producing the adjusted set  $\mathcal{T}_B^e$ .<sup>5</sup> The final *Clas-*

<sup>5</sup> The limited representation of some whistlers in the considered Silbo assortment, which constitutes the only data collection of its type and that will be described in Section 3.1, prevents the use of balancing strategies based on undersampling procedures.

*sification* stage then utilizes this balanced set to estimate the aforementioned function  $\hat{\Omega}$ .

During the inference phase, a query utterance  $q$  is drawn from a test set  $\mathcal{S}$ , which is disjoint from the training set  $\mathcal{T}$ —*i.e.*,  $\mathcal{T} \cap \mathcal{S} = \emptyset$ . The query  $q$  is processed by the *Feature extraction* phase, yielding the embedded representation  $q^e \in \mathbb{R}^f$ . Finally, the *Classification* method predicts the label of this query as  $\hat{y} = \hat{\Omega}(q^e)$ .

### 3 Experimental set-up

This section describes the data collection and evaluation protocol used to assess the proposed approach, as well as the embedding procedures, oversampling techniques, and classification strategies considered.

#### 3.1 Data assortment and evaluation protocol

We utilize the Silbo dataset compiled by Jakubiak [13], which represents the only existing assortment for the computational analysis of this language. This data collection comprises 529 whistled phrases recorded by 10 different practitioners, annotated at both word and sentence levels for transcription tasks. Table 1 provides further details on this collection in terms of the number of samples, total duration and average duration per sample.

Table 1: Details of the Silbo dataset compiled by Jakubiak [13] in terms of the number of samples, total duration and average duration per sample.

Number of Samples	Total Duration	Average Duration
529	1h 2m 3.3s	7.0s $\pm$ 2.9s

For the SI task, we exclusively consider the practitioner labels provided in the dataset as the target elements, disregarding all transcription-related annotations. Figure 2 illustrates the distribution of speaker identifiers in this dataset.

Although Jakubiak [13] proposed a partitioning scheme, it was specifically designed for transcription purposes and does not account for the distribution of practitioners across different partitions. To address this, our experiments adopt a 5-fold cross-validation scheme, stratified at the practitioner level. Within each fold, 10% of the training samples are set aside for validation. Note that this partitioning scheme results in a closed-set identification configuration in which the whistler to be identified is among the set of reference practitioners [33]. Open-set scenarios are posed as future work to be explored.

Regarding the evaluation protocol, we use the macro-average  $F_1$  score to measure the goodness of the proposal as it represents a standard figure of merit for

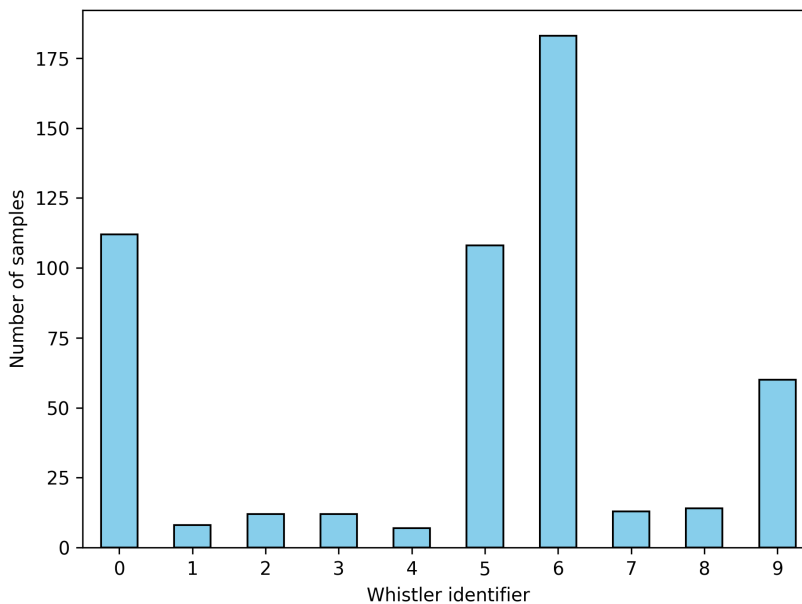


Fig. 2: Speaker identifier distribution of the Jakubiak dataset [13].

identification tasks [12] in contrast to other metrics more suitable for verification schemes such as Equal Error Rate [24]. The  $F_1$  score is defined as:

$$F_1 = \frac{1}{|\mathcal{C}|} \sum_{\forall c \in \mathcal{C}} \frac{2 \cdot TP_c}{2 \cdot TP_c + FP_c + FN_c} \quad (1)$$

where  $TP_c$ ,  $FP_c$ , and  $FN_c$  denote the True Positives, False Positives, and False Negatives for whistler identifier  $c \in \mathcal{C}$ , respectively.

### 3.2 Feature extraction

Given their competitive performance in the speech processing field, we adopt and compare three different feature extraction strategies for the identification task: the spectral-based mel-frequency cepstral coefficients (MFCC) commonly considered for speech analysis [8] together with two neural embedding schemes based on Speech Recognition models, *wav2vec 2.0* by Meta [1] and *Whisper* [31] by OpenAI. The remainder of this section provides a brief overview of these methods.

The *MFCC* descriptors constitute a set of low-level features which directly represent the spectral content of the utterance at hand. More precisely, this representation results from processing the Fourier Transform of the signal with a filter bank based on the perceptual Mel scale. We initially consider 20 filters (namely *Base*) and examine the influence of including both their first ( $Base + \Delta$ )

and second derivatives ( $Base + \Delta + \Delta^2$ ), as commonly done in SI with regular speech [33,24]. These cases result in embedding sizes of  $f = \{20, 40, 60\}$ , respectively.

The *wav2vec 2.0* scheme consists of a multi-layer convolutional neural network, whose output embeddings are processed by a Transformer network for transcription. This model is pre-trained and fine-tuned on 960 hours of the LibriSpeech dataset [29]. Note that we consider only the encoder component of the model as our objective is to obtain meaningful embedded representations of Silbo utterances. To evaluate its impact on the overall performance of our SI task, we experiment with different embedding sizes for the encoder, specifically  $f = \{64, 256, 1024, 4096\}$ .

The *Whisper* model is an end-to-end speech transcription system based on an encoder-decoder Transformer architecture. It is trained on 680,000 hours of multilingual and multitask supervised data collected from the Internet. As in the previous scheme, we exclusively consider the encoder stage of the scheme as it acts as a feature extractor for our task. In our experiments, we consider four versions of the model, which differ in the number of trainable parameters: *Tiny* (39M), *Base* (64M), *Small* (244M), and *Medium* (769M). The corresponding embedding sizes for these encoders are  $f = \{384, 512, 768, 1024\}$ , respectively.

### 3.3 Oversampling techniques

To mitigate the effects of the label imbalance present in the dataset, we employ three well-known oversampling strategies from the literature [26]: (i) Synthetic Minority Over-Sampling Technique (SMOTE) [4], (ii) Borderline SMOTE (B-SMOTE) [10], and (iii) Adaptive Synthetic Sampling (ADASYN) [11]. Since these methods require a feature-based representation of the data, the oversampling process is applied after the *Feature extraction* stage.

The SMOTE technique addresses class imbalance by generating synthetic samples in the regions of the  $\mathbb{R}^f$  feature space occupied by the minority classes. More precisely, the algorithm first selects a random sample from a minority class as well as one of its nearest neighbors in a random manner. A new synthetic sample is then generated by interpolating between the reference sample and the selected elements, with the new instance inheriting the reference sample’s label. This process is repeated for each class until a predefined balancing criterion is met (*e.g.*, ensuring all classes contain the same number of instances).

The B-SMOTE method extends the SMOTE algorithm by focusing on decision boundaries between classes. The oversampling process follows the same steps as SMOTE, with the key distinction that reference samples are specifically chosen from those lying on the decision frontiers—*i.e.*, instances predominantly surrounded by samples from the majority class.

The ADASYN algorithm differs from SMOTE-based strategies by employing an adaptive generation policy that prioritizes minority instances that are more difficult to classify, rather than uniformly sampling the minority classes. To achieve this, ADASYN uses a set of indicators to assess classification diffi-

culty in terms of label imbalance and generates synthetic samples accordingly to reduce these disparities.

### 3.4 Classification strategies

Regarding the *Classification* stage of the proposal, we examine five representative methods from the literature [7] with large application in the field of computational speech analysis, which are listed and described in the remainder of the section. Note that, for each classifier we assess different configuration parameters to optimize their recognition performance for the proposed SI task.

Based on its success in speech processing tasks, we examine the Gaussian Mixture Model (GMM) as a representative case of parametric learning [6]. For its use as a classifier, we fix the amount of gaussian functions in the mixture to the number of whistlers in the dataset. We initialize the centers of the distributions to those of the respective whistlers and optimize the rest of the parameters via Expectation-Maximization. We comparatively study the influence of the covariance function by comparing the case in which all components share the same general covariance matrix against that in which each component has its own single variance, respectively denoted as *tied* and *spherical* (SPH) in the rest of the work.

As a representative of the lazy learning paradigm, we consider the  $k$ -Nearest Neighbor ( $k$ NN) method, which classifies a given query based on the labels of the  $k$  elements that surround it in the feature space [35]. To assess the impact of the  $k$  hyperparameter on SI classification performance, we evaluate the method using  $k \in \{1, 3, 5\}$ .

In terms of neural networks, we explore the Multilayer Perceptron (MLP) [9] scheme as an example of this learning family. In this case, we analyze the effect of the optimization strategy by comparing classification performance when using the Limited-memory Broyden–Fletcher–Goldfarb–Shannon (LBFGS) algorithm versus the Stochastic Gradient Descent (SGD) method.

Regarding tree-based strategies, we evaluate the Random Forest (RaF) [15] method, which typically outperforms individual decision tree classifiers by leveraging an ensemble of such base classifiers to enhance robustness and reduce overfitting. To examine the influence of the number of trees in the ensemble, we test configurations with 100 and 500 trees.

Finally, given its competitive performance in related literature, we also include the Support Vector Machine (SVM) classifier in our study [17]. On this note, we compare two commonly used kernel functions: the polynomial one (Poly) and the Radial Basis Function (RBF).

## 4 Results

This section presents and discusses the results obtained for the SI task, following the experimental procedure outlined in Section 3. For clarity, the analysis is divided into two parts: (i) a comparative evaluation of feature extraction and

classification methods, focusing on base SI performance without applying balancing strategies, and (ii) an assessment of the impact of oversampling techniques on mitigating class imbalance to enhance the overall SI performance.

To ensure reproducibility and transparency, all developed code is publicly available at <https://github.com/jose-jvm/WhistlerIdentificationSilbo>. All experiments were conducted using Python (v. 3.11) with the *Hugging Face Transformers* (v. 4.46.3) [34], *librosa* [18], *scikit-learn* (v. 1.6.1) [30], and *imbalanced-learn* (v. 0.13.0) [16] libraries for the feature extraction, classification, and evaluation tasks. Finally, Table 2 summarizes the different experimental parameters assessed in the work.

Table 2: Summary of the experimental parameters considered in the experimentation categorized by those related to the embedding strategies, the classification methods, and the oversampling approaches.

Parameter	Value
<b><i>Feature extraction</i></b>	
MFCC	Base, Base + $\Delta$ , Base + $\Delta$ + $\Delta^2$
Wav2vec	64, 256, 1024, 4096
Whisper	Tiny, Base, Small, Medium
<b><i>Classification methods</i></b>	
Gaussian Mixture Models (GMM)	Tied, Spherical
$k$ -Nearest Neighbor ( $k$ NN)	1, 3, 5
Multi-Layer Perceptron (MLP)	Stochastic Gradient Descent Lim. Broyden–Fletcher–Goldfarb–Shannon
Random Forest (RaF)	100, 500
Support Vector Machine (SVM)	Polynomial, Radial Basis Function
<b><i>Oversampling approaches</i></b>	
Methods	Synthetic Minority Over-Sampling Technique Borderline-SMOTE Adaptive Synthetic Sampling

#### 4.1 Base identification performance

Table 3 presents the  $F_1$  score performance of the proposed SI scheme across different feature extraction and classification strategies. Note that this initial analysis does not incorporate any balancing methods.

Overall, MFCC-based representations consistently achieve the highest identification rates across all classifiers. The best result, an  $F_1$  score of 87.9%, is obtained using an SVM classifier with a polynomial kernel and MFCC features augmented with first-order derivatives (Base +  $\Delta$ ). This finding aligns with the

Table 3: Average test results for the 5-fold cross-validation scheme in terms of  $F_1$  (%) for the classifiers evaluated with respect to the embedding strategy without oversampling. Bold values highlight the best result for each classification scheme and embedding method, while underlined values indicate the best overall result per classifier. Feature sizes ( $f$ ) are provided for comparison.

	GMM		$k$ NN			MLP		RaF		SVM	
	SPH	Tied	1	3	5	LBFGS	SGD	100	500	Poly	RBF
<b>MFCC</b>											
Base (20)	24.5	72.7	75.2	70.5	68.8	80.5	23.4	<b>68.8</b>	66.7	85.8	22.2
Base + $\Delta$ (40)	22.3	<b>79.3</b>	<b>76.2</b>	70.9	70.3	<b>82.5</b>	7.7	64.9	66.1	<b>87.9</b>	22.1
Base + $\Delta$ + $\Delta^2$ (60)	19.1	78.0	75.6	67.7	67.4	78.8	16.9	66.2	68.3	86.3	22.2
<b>Wav2vec</b>											
64	8.0	9.6	12.2	12.0	<b>14.0</b>	10.7	12.6	10.2	10.3	12.5	10.7
256	7.5	<b>11.3</b>	12.3	12.2	13.9	11.7	12.0	11.9	11.8	<b>13.8</b>	13.5
1024	6.5	3.6	11.3	10.4	10.1	12.3	11.9	13.2	<b>13.7</b>	12.7	<b>13.8</b>
4096	6.8	1.9	12.4	10.1	8.7	16.0	14.7	12.0	11.6	13.0	11.0
<b>Whisper</b>											
Tiny (384)	7.1	49.5	39.6	37.9	29.2	<b>57.1</b>	7.8	26.3	23.4	83.7	11.7
Base (512)	10.6	<b>65.4</b>	<b>47.4</b>	38.7	36.9	45.2	5.1	34.7	<b>36.2</b>	<b>83.8</b>	11.0
Small (768)	7.7	23.0	26.3	20.7	19.3	31.2	5.1	21.7	21.4	71.3	9.1
Medium (1024)	12.8	59.0	34.3	35.4	27.5	34.4	4.5	32.2	30.2	74.5	7.4

speaker recognition literature, confirming the effectiveness of MFCCs for distinguishing individual whistlers in Silbo, likely due to the prominent spectral patterns in whistled speech.

Focusing on the particular MFCC configurations, it is observed that incorporating first-order derivatives (Base +  $\Delta$ ) generally improves performance compared to using only the base coefficients, suggesting that the inclusion of dynamic spectral information benefits the discrimination between whistlers. Interestingly, the best identification rates are achieved with the first-order configuration (Base +  $\Delta$ ) across all classifiers except for RaF, which performs best with the base coefficients alone.

In contrast, Wav2vec embeddings yield the lowest performance across classifiers, with  $F_1$  scores typically below 15%. This suggests that features learned from spoken speech do not generalize well to whistled signals within the SI context. Conversely, Whisper-based embeddings show improved performance, achieving  $F_1$  scores up to 83.8% with the SVM classifier. This indicates that large-scale multilingual pre-training confers better generalization to non-verbal speech modalities such as whistling. Among the Whisper configurations, model complexity plays an important role, with the best results obtained using the less complex Tiny and Base models.

Finally, the choice of classification scheme plays a critical role in the SI task, with different methods showing varying sensitivities to parameter tuning. The SVM classifier with a polynomial kernel consistently delivers the highest perfor-

mance across feature types, highlighting its capacity to generalize effectively in this context. Tree-based methods (RaF) and  $k$ NN classifiers also provide stable yet slightly lower performance with minimal tuning. In contrast, methods such as MLP and GMM can achieve competitive results but require careful configuration to reach optimal performance.

In summary, the results obtained establish a strong baseline for speaker identification in Silbo, demonstrating that classical spectral features based on cepstral principles combined with well-tuned classifiers can achieve high recognition rates even in label-imbalance cases.

## 4.2 Oversampling strategies for data balancing

Following the initial analysis, this section examines the impact of oversampling techniques on addressing label imbalance within the Silbo dataset. For conciseness, we focus on the best-performing classifier configurations identified in Section 4.1: GMM with Tied covariance,  $k$ NN with  $k = 1$ , MLP with the LBFGS optimizer, RaF with 500 trees, and SVM with the Poly kernel. Additionally, we restrict the analysis to the MFCC and Whisper feature extraction methods under the configurations yielding the highest performance for each classifier.

Table 4 presents the  $F_1$  scores obtained for the proposed SI scheme when applying different oversampling strategies. The None case—*i.e.*, no balancing method is applied—serves as the baseline for comparison.

Table 4: Average test results for the 5-fold cross-validation scheme in terms of  $F_1$  (%) for the best classification configurations from Table 3, evaluated across different oversampling strategies. Bold values indicate the best result per embedding method, classification scheme, and oversampling strategy, while underlined values highlight the best overall score per embedding method.

	Oversampling method			
	None	SMOTE	B-SMOTE	ADASYN
<b><i>MFCC</i></b>				
GMM	79.3	78.5	79.3	81.7
$k$ NN	76.2	79.6	78.2	79.0
MLP	82.5	72.1	81.3	69.0
RaF	68.8	73.2	73.3	73.5
SVM	<b>87.9</b>	<b>86.2</b>	<b>85.8</b>	<b>86.2</b>
<b><i>Whisper</i></b>				
GMM	65.4	72.2	56.2	62.7
$k$ NN	47.4	53.9	51.4	54.0
MLP	57.1	68.7	70.2	66.3
RaF	36.2	58.5	53.3	59.2
SVM	<b>83.8</b>	<b>83.7</b>	<b>83.5</b>	<b>85.2</b>

The results indicate that applying oversampling generally improves classification performance, although the magnitude of enhancement depends on both the feature representation and the baseline performance without balancing (*i.e.*, the None case). For MFCC features, increases in the figure of merit are observed primarily for the GMM,  $k$ NN, and RaF classifiers, while MLP and SVM do not benefit to the point of achieving their best performance without oversampling. For instance, RaF improves from 68.8% to 73.5%  $F_1$  with ADASYN, and  $k$ NN increases from 76.2% to 79.6% with SMOTE, while SVM maintains its highest performance (87.9%  $F_1$ ) without balancing, with only marginal changes across oversampling methods.

In the case of Whisper-based features, oversampling consistently yields improvements across  $k$ NN, MLP, and RaF, indicating that balancing is particularly beneficial when the feature representation alone does not ensure sufficient class discrimination. Notably, RaF improves from 36.2% to 59.2%  $F_1$  with ADASYN, and MLP increases from 57.1% to 70.2% with B-SMOTE. SVM, while already performing well with Whisper features, shows a slight improvement from 83.8% to 85.2%  $F_1$  with ADASYN. GMM, in contrast, only benefits with SMOTE, improving from 65.4% to 72.2%, but exhibits lower results with other oversampling strategies.

Among the oversampling methods evaluated, ADASYN consistently provides the highest or near-highest performance across multiple configurations. This aligns with its adaptive sampling policy, which prioritizes generating synthetic examples in regions where minority classes are harder to classify, unlike SMOTE-based methods that distribute synthetic samples more uniformly.

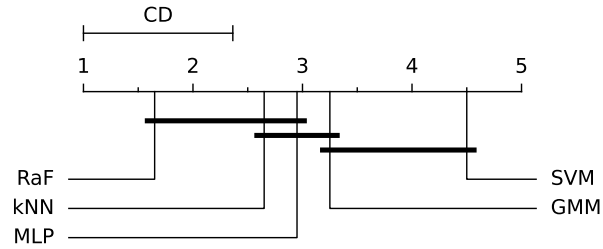
Within the SMOTE family, no clear overall advantage emerges between SMOTE and B-SMOTE, with their relative performance highly depending on the classifier-feature configuration. For example, B-SMOTE slightly outperforms SMOTE with MLP on Whisper features, while the inverse is true for  $k$ NN.

Overall, the best results across all configurations are achieved using the SVM classifier with MFCC features, maintaining an  $F_1$  score of 87.9% without requiring oversampling. For Whisper features, the highest  $F_1$  score (85.2%) is achieved with SVM using ADASYN, illustrating the potential of oversampling to close the performance gap between Whisper and MFCC representations under optimal configurations.

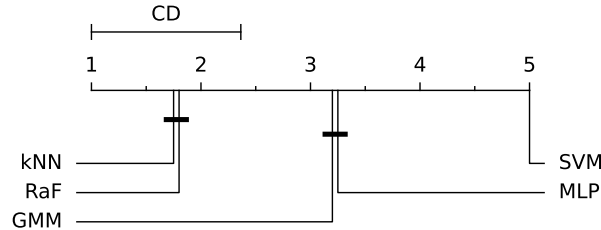
To conclude the analysis, we conduct a Nemenyi post-hoc test [5] to assess the statistical significance of differences among classifiers. Figure 3 shows the results for both MFCC and Whisper features, considering a significance threshold of  $p < 0.05$ .

As it can be observed, the analysis confirms the similar performance of the  $k$ NN, RaF, and MLP classifiers under many configurations, while the SVM consistently ranks highest. This fact reinforces its superior generalization capacity for SI in Silbo, as previously observed.

Finally, we conclude that while oversampling can remarkably enhance performance in label-imbalanced and low-separability settings, high-performing configurations such as MFCC combined with SVM do achieve strong results without



(a) Feature representation using MFCC.



(b) Feature representation using the Whisper encoder.

Fig. 3: Results of the Nemenyi post-hoc test assessing the relative improvement across the different classification strategies for the MFCC and Whisper representations depicting the most competitive results. The Critical Distance (CD) represents the minimum difference in the significance score to consider a pair of classifiers as statistically different.

requiring data balancing. This establishes a solid benchmark for future computational studies on whistled language SI under data-scarce scenarios.

## 5 Conclusions and future work

The whistled Spanish of the Canary Islands, or Silbo, is one of the world’s most representative whistled languages and has long been the focus of a number of linguistic studies. As the most intensively studied form of whistled speech worldwide, Silbo has yielded insights into ethnological, linguistic, and cognitive phenomena, but computational methodologies for advanced, automated analysis remain largely unexplored.

This work presents the first approach to Speaker Identification (SI)—*i.e.*, the process of determining the speaker of a given utterance via computational means—tailored to Silbo in a closed-set disposition. Leveraging a combination of standard spectral-based feature extraction methods, embeddings from pre-trained Speech Recognition models, and class-balancing techniques, we prove that automatic SI for Silbo can successfully be carried out with competitive performance rates under data-scarce conditions.

Extensive experiments conducted on the Jakubiak dataset [13]—the only resource specifically compiled for computational analysis of Silbo—show that MFCC-based features, particularly when combined with an SVM classifier using a polynomial kernel, deliver the best performance, achieving an  $F_1$  score of 87.9%. Whisper-based embeddings also demonstrate competitive results, reaching up to 85.2%  $F_1$  with oversampling, although their effectiveness is highly dependent on the classification strategy considered. In contrast, embeddings derived from Wav2vec exhibit limited applicability in this context, highlighting the challenges of directly transferring features learned from monolingual spoken speech to whistled modalities.

Our findings highlight that, while oversampling techniques can remarkably enhance performance in scenarios with lower class separability (e.g., Whisper-based features), high-performing configurations such as MFCC with SVM can achieve strong results without requiring data balancing. This establishes a robust baseline for computational SI in Silbo, demonstrating the potential of standard feature representations and classifiers in addressing this underexplored task. However, despite these promising results, the scarcity of whistled data severely hinders progress in this field, which highlights the need for larger datasets.

Based on the above, future work will focus on expanding existing data collection to include a larger and more diverse collection of practitioners, whistling styles, and recording conditions. Other aspects to be explored comprise the adaptation or fine-tuning of neural-based embedding models for extracting adequate representations for whistled speech, or the use of data-efficient approaches such as Siamese networks for low-resource SI scenarios. Additionally, the integration of multimodal representations, combining spectral, temporal, and possibly articulatory features, may further enhance system robustness and accuracy. Extending this work to open-set SI, speaker verification, and speaker diarization tasks in real-world whistled speech recordings is also a promising direction for future exploration.

All in all, our work lays a solid foundation for the computational analysis of Silbo, showing the feasibility and effectiveness of SI for whistled speech and bridging the gap between traditional linguistic research and modern machine learning techniques within the digital humanities.

**Acknowledgments.** This work was partially funded by the Generalitat Valenciana through project CIGE/2023/216.

## References

1. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems* **33**, 12449–12460 (2020)
2. Bai, Z., Zhang, X.L.: Speaker recognition based on deep learning: An overview. *Neural Networks* **140**, 65–99 (2021)
3. Busnel, R.G., Classe, A.: Whistled languages, vol. 13. Springer Science & Business Media (2013)

4. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* **16**, 321–357 (2002)
5. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research* **7**, 1–30 (2006)
6. Dhanjal, A.S., Singh, W.: A comprehensive survey on automatic speech recognition using neural networks. *Multimedia Tools and Applications* **83**(8), 23367–23412 (2024)
7. Duda, R.O., Hart, P.E., et al.: *Pattern classification*. John Wiley & Sons (2006)
8. Ganchev, T., Fakotakis, N., Kokkinakis, G.: Comparative evaluation of various mfcc implementations on the speaker verification task. In: *Proceedings of the International Conference of Speech and Computer (SPECOM)*. vol. 1, pp. 191–194 (2005)
9. Han, B., Chen, Z., Liu, B., Qian, Y.: Mlp-svnet: A multi-layer perceptrons based network for speaker verification. In: *International Conference on Acoustics, Speech and Signal Processing*. pp. 7522–7526 (2022)
10. Han, H., Wang, W.Y., Mao, B.H.: Borderline-smote: a new over-sampling method in imbalanced data sets learning. In: *International Conference on Intelligent Computing*. pp. 878–887 (2005)
11. He, H., Bai, Y., Garcia, E.A., Li, S.: Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: *International Joint Conference on Neural Networks*. pp. 1322–1328 (2008)
12. Jahangir, R., Teh, Y.W., Nweke, H.F., Mujtaba, G., Al-Garadi, M.A., Ali, I.: Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges. *Expert Systems with Applications* **171**, 114591 (2021)
13. Jakubiak, A.: Whistle-to-text: Automatic recognition of the silbo gomero whistled language. In: *Proceedings of the 24th INTERSPEECH Conference*. pp. 3402–3406 (2023)
14. Johansson, A.T., White, P.R.: An adaptive filter-based method for robust, automatic detection and frequency estimation of whistles. *The Journal of the Acoustical Society of America* **130**(2), 893–903 (2011)
15. Karthikeyan, V., Suja Priyadharsini, S.: Adaptive boosted random forest-support vector machine based classification scheme for speaker identification. *Applied Soft Computing* **131**, 109826 (2022)
16. Lemaître, G., Nogueira, F., Aridas, C.K.: Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research* **18**(17), 1–5 (2017)
17. Malik, M., Malik, M.K., Mehmood, K., Makhdoom, I.: Automatic speech recognition: a survey. *Multimedia Tools and Applications* **80**, 9411–9457 (2021)
18. McFee, B., Raffel, C., Liang, D., Ellis, D.P., McVicar, M., Battenberg, E., Nieto, O.: librosa: Audio and music signal analysis in python. *SciPy* **2015**, 18–24 (2015)
19. Meyer, J.: Whistled languages. *A Worldwide Inquiry on Human Whistled Speech* (2015)
20. Meyer, J.: Environmental and linguistic typology of whistled languages. *Annual Review of Linguistics* **7**(1), 493–510 (2021)
21. Meyer, J., Díaz Reyes, D.: Geolingüística de los lenguajes silbados del mundo, con un enfoque en el español silbado. *Géolingüistique* (17), 99–124 (2017)
22. Meyer, J., Magnasco, M.O., Reiss, D.: The relevance of human whistled languages for the analysis and decoding of dolphin communication. *Frontiers in Psychology* **12**, 689501 (2021)

23. Meyer, J., Rolland, V., Socas, T., Díaz, D.: A sentence comprehension test with whistled spanish experts. In: ExLing Conferences. pp. 65–68 (2024)
24. Mittal, A., Dua, M.: Automatic speaker verification systems and spoof detection techniques: review and analysis. *International Journal of Speech Technology* **25**(1), 105–134 (2022)
25. Mohd Hanifa, R., Isa, K., Mohamad, S.: A review on speaker recognition: Technology and challenges. *Computers & Electrical Engineering* **90**, 107005 (2021)
26. Mujahid, M., Kima, E., Rustam, F., Villar, M.G., Alvarado, E.S., De La Torre Diez, I., Ashraf, I.: Data oversampling and imbalanced datasets: An investigation of performance for machine learning and feature engineering. *Journal of Big Data* **11**(1), 87 (2024)
27. Ngoc, A.T., Meyer, J., Meunier, F.: The effect of musical expertise on whistled vowel identification. *Speech Communication* **159**, 103058 (2024)
28. O’Brien, B., Marczyk, A.: A spectrotemporal modulation application for distinguishing modal and whistled speech. *International Journal of Speech Technology* pp. 1–8 (2025)
29. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: Librispeech: an asr corpus based on public domain audio books. In: *International Conference on Acoustics, Speech and Signal Processing*. pp. 5206–5210 (2015)
30. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011)
31. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: *Proceedings of the 40th International Conference on Machine Learning*. pp. 28492–28518 (2023)
32. Tapiador, F.J.: Heritage: A treasure chest. *The Geography of Spain: A Complete Synthesis* pp. 405–419 (2020)
33. Tirumala, S.S., Shahamiri, S.R., Garhwal, A.S., Wang, R.: Speaker identification features extraction methods: A systematic review. *Expert Systems with Applications* **90**, 250–271 (2017)
34. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al.: Transformers: State-of-the-art natural language processing. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. pp. 38–45 (2020)
35. Yerramreddy, D.R., Marasani, J., Gowtham, P.S.V., Yashwanth, S., SS, P., et al.: Speaker identification using mfcc feature extraction: A comparative study using gmm, cnn, rnn, knn and random forest classifier. In: *International Conference on Trends in Electrical, Electronics, and Computer Engineering*. pp. 287–292 (2023)