# Phoneme Analysis for Multiple Languages with Fuzzy-Based Speaker Identification

DE LIMA, Thales Aguiar and DA COSTA ABREU, Marjory <http://orcid.org/0000-0001-7461-7570>

Received: 14 March 2022 | Revised: 24 April 2022 | Accepted: 25 April 2022

IET Biometrics

DOI: 10.1049/bme2.12078

CASE STUDY

# Phoneme analysis for multiple languages with fuzzy-based speaker identification

**Thales Aguiar de Lima**[1] | **Márjory Cristiany Da-Costa Abreu**[2]

[1]Department of Informatics and Applied Mathematics, Federal University of Rio Grande do Norte, Natal, Brazil

[2]Department of Computing, Sheffield Hallam University, Sheffield, UK

**Correspondence**

Thales Aguiar de Lima, Department of Informatics and Applied Mathematics, Federal University of Rio Grande do Norte, 59078-900, Natal, RN 1524, Brazil.
Email: thales.aguiar.016@ufrn.edu.br

**Funding information**

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Grant/Award Number: 001

**Abstract**

Most voice biometric systems are dependent on the language of the user. However, if the idea is to create an all-inclusive and reliable system that uses speech as its input, then they should be able to recognise people regardless of language or accent. Thus, this paper investigates the effects of languages on speaker identification systems and the phonetic impact on their performance. The experiments are performed using three widely spoken languages which are Portuguese, English, and Chinese. The Mel-Frequency Cepstrum Coefficients and its Deltas are extracted from those languages. Also, this paper expands the research study of fuzzy models in the speaker recognition field, using a Fuzzy C-Means and Fuzzy k-Nearest Neighbours and comparing them with k-Nearest Neighbours and Support Vector Machines. Results with more languages decreases the accuracy from 92% to 85.59%, but further investigation suggests it is caused by the number of classes. A phonetic investigation finds no relation between the phonemes and the results. Finally, fuzzy methods offer more flexibility and in some cases, even better results compared to their crisp version. Therefore, the biometric system presented here is not affected by multilingual environments.

**KEYWORDS**

artificial intelligence, biometrics, Portuguese, speaker identification, voice

## 1 | INTRODUCTION

Speech exists with the main reason to enable communication between humans. This communication translates into a sequence-dependent and rule-based system: *a language*. To talk with each other, humans use a complex system to produce the voice signal. Starting at the lungs, though the trachea, stimulating vocal cords and the larynx tube, using the pharynx cavity, the tongue, vellum, mouth, and nasal cavity to produce sound. This procedure is detailed in Ref. [1].

Speaker Identification (SI) is a biometric branch of the Automatic Speech Recognition field. It can be defined as 'deciding if a speaker is a specific person or is among a group of persons.' without prior identity claim [2]. It is popularly used for database search in criminal records [3]. For instance, ALIZE is a popular framework [4]. Thus, this biometric focusses on *identity recognition*. This problem can be further specified as *open-set* when the speaker is not enrolled in the system and as *closed-set* when everyone is registered [5]. Furthermore, some systems rely on previous knowledge of what is said, that is, a type of passphrase. Those are classified as *text-dependent*, in contrast to *text-independent* when the user can speak anything [6]. This paper explores *closed-set text-independent speaker identification* systems. More specifically, this paper proposes a method to verify how different languages (on structure, accent, and ancestry): Portuguese (BP), English (EN), and Chinese (CN), can affect fuzzy and typical classifiers for those systems.

### 1.1 | Literature review

The literature presents a wide range of experiments for this biometry. To represent a speaker, the Mel-Frequency Cepstrum

Coefficients (MFCC) are widely adopted [7–10] as voice-print, even though the state of the art has shifted from it to *i*-vectors [9, 11] and then towards *x*-vectors [12]. Also, there are some variants for those representations [13] and few using fuzzy information theory [14, 15].

Besides biometric features, classification has also improved for SI. For long, the Gaussian Mixture Model combined with Universal Background Model [6] and Hidden-Markov Models [16] dominated the field. However, other methods such as vector quantisation [17, 18] have their spots. A little research is made for fuzzy classification [8, 19–22], but most are quite dubious when describing their methods for both models and data. Furthermore, most recent research has converged to Neural Network variants, such as Deep Neural Networks [11, 23, 24], Convolutional Neural Networks [25], and others [26, 27].

Meanwhile, the SI community has always speculated the impact of language on those systems [28]. In fact, some studies investigate this topic [29, 30], but they usually employ languages with common ancestry such as English and German, or even accent variations. Other limitations on these studies are the use of a small dataset, not providing a better description of how to split the dataset or any statistical tests performed. Also, research for fuzzy models is scarce, even though they have provided decent results. Moreover, only one work [31] has considered BP in its open-set classification. The low occurrence of this language is due to its lack of resource for creating speech technologies [32].

Therefore, the main contributions of this paper are as follows:

- Propose a method to verify how different languages, without ancestry, can affect closed-set text-independent speaker identification;
- Expand the research with Brazilian Portuguese;
- A phonetic analysis to investigate how the sounds can affect those systems;
- Provide a detailed discussion and methodology for Fuzzy models applied to speaker identification.

The rest of the paper describes the datasets, the data processing, and feature extraction methodology of this work (Section 2). Next, there is a brief description of the classification methods (Section 3). Then, it introduces the evaluation methodology with the hyperparameter space of the models (Section 4). Next, the results are presented (Section 5), including a phonetic analysis and followed by an extensive discussion (Section 6). The paper ends with the final remarks, presenting some limitations and a summary of the outcomes of this research study (Section 7).

## 2 | DATA AND PRE-PROCESSING

This section presents the building-steps of the multilingual dataset. Besides, it also describes the pre-processing methods (data cleaning, under-sampling, and reduction) and finally the process to extract features.

### 2.1 | Multilingual data

This research study uses three distinct datasets on three different languages: DARPA-TIMIT [33], LapsBenchmark16k (LAPSBM16K) [34], and AISHELL-1 [35]. They provide data on EN, BP, and CN, respectively. The choice of the datasets is based on their equal sampling rate, besides all of them being public and free. Other multilingual corpus is the NIST SRE datasets [36], Call My Net Corpus [37], and more. However, most of them have a price and are under the Linguistic Data Consortium (LDC), which puts them over the budget of this research study. Each dataset is better detailed in their respective references; therefore, the following is merely a brief description of their main characteristic.

The DARPA-TIMIT is a free version of TIMIT, which has to be purchased at the LDC. The prompts on the corpora are scripted, and every speaker has a total of 10 samples. The audio in this dataset has 2.9s $\pm$ 0.8s of duration. The BP data from LAPSBM16K has 20 audio samples per speaker, while their durations are about 4.6s $\pm$ 0.8s. Finally, the AISHELL-1 provides a substantial amount of CN speech with at least 300 samples per speaker and an approximated duration of 4.6s $\pm$ 1.3s.

Table 1 summarises the main characteristics of the datasets. The gender distribution from BP and EN are not good compared to that of CN. However, since our goal is to investigate multilingual speech technologies, then overall the gender is almost evenly distributed. The number of recordings for each data is also quite different. However, these characteristics are balanced by under-sampling them when needed. Furthermore, the most important is that all recordings have the same sampling rate of 16 kHz/16bit, which ensures the same resolution across languages. During review, the representativeness of the data was questioned. However, DARPA-TIMIT has much more depth than described here, and the Chinese dataset is large. The Brazilian Portuguese dataset is small compared to the others; however, there are not many resources for this language.

### 2.2 | Pre-processing the datasets

The first step of our work was to make sure all the dataset distributions were as similar as possible. Two main characteristics kept from the data were gender and number of speakers. Since LAPSBM16K had the least number of classes, both EN

**TABLE 1** Summary of the datasets

| Dataset | Size | #Speakers | Gender (M/F) | Lang | Source |
|---|---|---|---|---|---|
| DARPA-TIMIT | 6300 | 630 | 70%/30% | EN | [33] |
| LAPSBM16K | 700 | 35 | 72%/28% | BP | [35] |
| AISHELL-1 | 141,200 | 400 | 47%/53% | CN | [34] |
| Total | 148,200 | 1065 | 48%/52% | — | — |

Abbreviations: BP, Portuguese; CN, Chinese; EN, English.

and CN had to be adjusted to obtain a fairer experimental setup.

Also, DARPA-TIMIT had the least samples per speaker. Thus, an under-sampling is applied on AISHELL-1 and LAPSBM16K, making them have 10 samples for each speaker through a Roulette algorithm. Furthermore, the total number of classes for BP is 35, way less than other languages. Therefore, some English speaker had to be cut-off from experiments resulting into 34 speakers. The development subset with 40 speakers was used for the CN dataset. Figure 1 gives an overall comparison between the original and experimental multilingual dataset.

### 2.3 | Extracting speaker features

To prepare the data for the SI task, the signals pass through an energy-based voice activity detection. The experiments use the first 13 coefficients from 40 extracted MFCC, excluding the $0th$, using frames of 25 ms length with 10 ms stride, a Hamming Window function as well as a 512-point Fast Fourier Transform, besides, 40 triangular filters spamming from 300 to 3400 Hz. Then, calculating the Deltas and double Deltas and appending the logarithm energy, a 40-dimensional MFCC feature vector is created. Finally, a cepstral mean subtraction is applied to the features before training/testing to remove channel and recording variations [6].

This section briefly introduced our three datasets: DARPA-TIMIT, LAPSBM16K, and AISHELL-1. Also, it introduced the pre-process of the data and the process of feature extraction.

## 3 | METHODS

Four classification models are used in this research study. There are two fuzzy methods: *Fuzzy C-Means* (FCM) and *Fuzzy k-Nearest Neighbours* (FkNN), and two traditional methods: *k-Support Vector Machines ours* (kNN) and *Support Vector Machines* (SVM). Since the traditional methods are well known by the research community, this section focusses on describing the Fuzzy classifiers.

### 3.1 | Fuzzy C-Means

The Fuzzy C-Means [38] and its improved version [39] works by minimising the weighted distance of each sample to every cluster centre. The weights correspond to the membership matrix $U = u_{ij}$. Thus, given a fuzziness factor $m$, a sample $x_i$ has a membership degree of $u_{ij}$ for cluster $j$. Therefore, every cluster works as a fuzzy subset. Those subsets have a fuzziness degree $m > 1$, which affects the compatibilities of data samples according to

$$u_{ij} = \sum_{k=1}^{C} \left( \frac{|x_i - c_j|}{|x_i - c_k|} \right)^{\frac{-1}{m-1}} \tag{1}$$

### 3.2 | Fuzzy *k*-Nearest Neighbours

Another fuzzification of crisp methods is the Fuzzy $k$-Nearest Neighbours. In this paper, this model is based on the work of Ref. [40]. First, the model creates a membership matrix $U_{ij}$ with the labelled data using a kNN [40]. With $n_j$ neighbours of $x$ that belong to class $i$, $L \neq_n K$ is the total number of neighbours. Then, it classifies new data by finding the closest $k$ vectors, scaling the fuzzified distances by $u_{ij}$ and normalising by the sum of distances.

$$u_{ij}(x) = \begin{cases} 0.51 + \frac{n_j}{L} \cdot 0.49, & i = j \\ \frac{n_j}{L} \cdot 0.49 & i \neq j \end{cases} \tag{2}$$

$$u_i(x') = \frac{\sum_{j}^{K} u_{ij} |x' - x_j|^{\frac{-2}{m-1}}}{\sum_{j}^{K} |x' - x_j|^{\frac{-2}{m-1}}} \tag{3}$$

## 4 | EXPERIMENTAL SETUP

The following steps were executed on a system with an AMD Ryzen-5 1600 Six-Core Processor, Dual Channel 2 × 8 GiB DIMM DDR4 2400 MHz, SSD Kingston A1000 NVMe R1500 Mb/s, and W500 Mb/s, [MSI] Radeon RX580 8G OC, 64-bit Pop!_OS 20.04 with Gnome 3.36.2.

The method consists of three stages: (i) the monolingual, where a baseline accuracy for BP (after pre-processing) is defined, and the model with the best performance is used for further experiments; (ii) then, in the multilingual, the monolingual data is augmented with EN speakers, and then with CN to showcase the SI performance when adding more languages; (iii) finally, the size constraint, creates smaller multilingual datasets to investigate if the number of classes is affecting the results.

A total of 30 experiments with 34 speakers from all languages are performed to verify any bias towards number of classes. For that, 1/3 of data from each language (considering Figure 1 left) is selected, this time ignoring other characteristics from data.

The description below presents every set of configurations for each parameter used during the fine-tuning. This procedure was performed with a stratified 3-fold cross validation and a grid-search with the hyperparameters below. The stratified version can preserve the class distribution, while a 3-fold guarantees a decent trade-off between the amount of training and test samples. The kNN and SVM are from the SKLearn library [41], while FCM and FkNN are implemented by the authors using python and is available at GitHub[1].

**FCM** have $m$ varying in {1.5, 2, 2.5, 3}. The number of clusters is fixed at the number of classes, and the stop criterion

[1] https://github.com/thalesaguiar21/Fuzzy

used is a fixed tolerance of 0.2. Tests used three distances: Manhattan, Euclidean and Minkowski.

**FkNN** have $K \in \{2, 4, \ldots, 12\}$. Distance metrics are Manhattan, Euclidean, and Minkowski for $m \in \{1.5, 2, 2.5, 3\}$ and $L$ in (3) is fixed to 16.

**kNN** have $K \in \{2, 4, \ldots, 12\}$. Distance metrics are Manhattan, Euclidea, Minkowski, and DTW.

**SVM** has $C$ and $\gamma \in \{0.0001, 0.001, 0.01, 0.1, 1, 10\}$. Kernels tested are polynomial, radial basis function (RBF), and sigmoid. When using the polynomial kernel, the degrees {1, 2, 3, 4, 5} are verified. Thus, the linear kernel is verified as well.

## 5 | EXPERIMENTAL RESULTS

The results are divided into sections with respect to each experiment performed in this paper.

### 5.1 | Monolingual

Results for Fuzzy clustering reaches a maximum accuracy score of 32.57% ± 4.88% (Figure 2). Setting the fuzziness $m = 1.5$ and the metric to Euclidean produces the best results, while cluster fuzziness has no significant improvement.

FkNN, on the other hand, achieves 87.42% ± 4.1% accuracy (Figure 3). This score is obtained when using the Euclidean similarity with 2 neighbours and $m = 2$. Therefore, it represents an absolute 54.85% improvement over FCM. Besides, increasing $m$ above 2 makes it decrease the classification score. Also, notice that it obtained its best scores when using a small $k$.

For kNN, the best value is 86% ± 3.14% (Figure 4), a 1.42% attenuation compared to FkNN. This score is achieved by Euclidean, Minkowski, and Manhattan metrics using $k = 6$. The last metric is the best, as it has the same performance with better generalisation.

With respect to the SVM, almost every kernel achieves decent accuracy values. Both Sigmoid and RBF get the highest score of 92.29% ± 4.8%. They obtain this result when $\gamma = 10^{-3}$ and $C = 10$, while the Linear kernel is close by with 92% ± 4.2% accuracy.

Since there is a tiny difference between these results, it was necessary to compare these kernels with more details. In short, the Linear kernel has a 7.7% better performance per time (Table 3), lower $\sigma^2$ (Table 2), smaller test duration (Table 2), and lower $C$ (Table 3), while losing in absolute accuracy and $\gamma$. Since using a $C = 10$ can lead to a non-generic model, the Linear SVM configuration (SVML1-C01G01) is better suited for this problem. SVML1-C01G01 shows 4.58% improvement over FkNN. Table 2 presents the best results of each kernel. Finally, Figure 5 presents the Linear SVM results in more detail.

### 5.2 | Multilingual

Now, SVML1-C01G01 is submitted to experiments with BP + EN and BP + EN + CN speakers. Results are presented in Figure 6.

Results shown on the 2-language experiments have a similar behaviour from monolingual SVM tests regarding $C$ with no effect on accuracy for $\gamma \leq 0.01$. The best configuration on this dataset is still $C = 0.01$ and $\gamma = 10^{-1}$, achieving 87.97% ± 2.56% of accuracy and therefore a 4.03% decrease compared to the

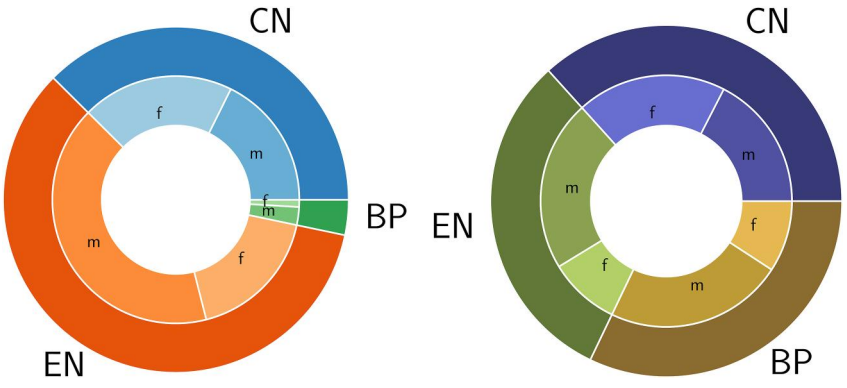


**FIGURE 1** Original (left) and experimental (right) data distribution for number of speakers and gender

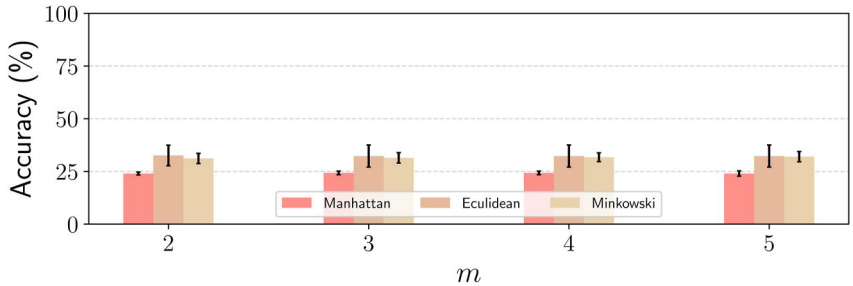


**FIGURE 2** Results for FCM in Brazilian Portuguese. FCM, fuzzy C-means

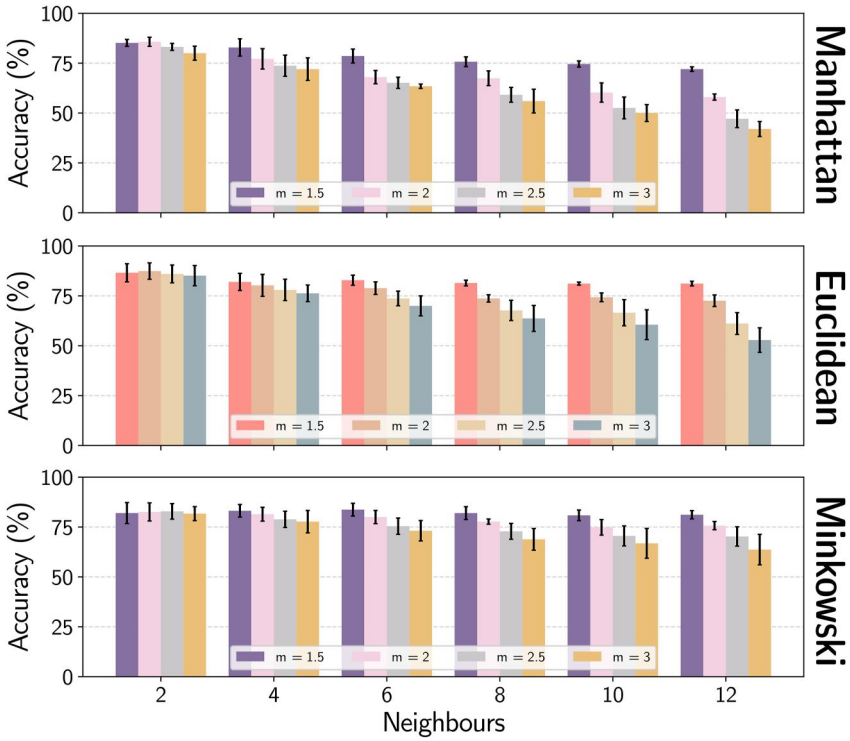


**FIGURE 3** Results for FkNN in Brazilian Portuguese. FkNN, Fuzzy k-Nearest Neighbours

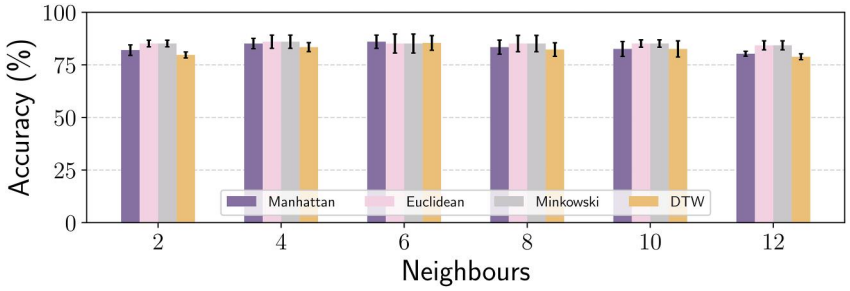


**FIGURE 4** kNN results for Brazilian Portuguese. kNN, k-Nearest Neighbours

**TABLE 2** Monolingual results of SVM

| Model | $\overline{acc}$ (%) | $\sigma^2$ (%) |
|---|---|---|
| SVM-Poly1-C01-G01 | 92.00 | 4.20 |
| SVM-Poly2-C1-G0001 | 79.15 | 3.14 |
| SVM-Poly3-C1-G0001 | 81.15 | 4.50 |
| SVM-Poly4-C01-G001 | 69.62 | 5.10 |
| SVM-Poly5-C01-G01 | 71.44 | 7.20 |
| SVM-RBF-C10-G0001 | 92.29 | 4.80 |
| SVM-Sig-C10-G0001 | 92.29 | 4.80 |

Abbreviation: SVM, Support Vector Machines.

**TABLE 3** Accuracy over time for classifiers

| Model | Train (ps) | Test (ps) | ACC/ps (%) |
|---|---|---|---|
| FCM | 300.00 | 75.00 | 00.80 |
| FkNN | 267.00 | 128.00 | 00.32 |
| kNN | 0.05 | 0.58 | 14.90 |
| Poly1-SVM | 7.02 | 2.21 | 41.40 |
| Sig-SVM | 8.42 | 2.74 | 33.70 |
| RBF-SVM | 8.42 | 2.74 | 33.70 |

Abbreviations: FCM, Fuzzy C-Means; FkNN, Fuzzy k-Nearest Neighbours; kNN, k-Nearest Neighbours; RBF, radial basis function; SVM, Support Vector Machines.

monolingual experiments. Also, the number of mistakes of BP speakers by EN is small when compared to the opposite, as presented in Figure 7. Since the features were carefully extracted and processed to remove any bias from languages, recording procedures, or errors added while transforming the signal; these mistakes are unlikely originated from those sources.

Next, adding CN as a third language to the dataset, results into 85.59% ± 1.32% accuracy and 2.38% decrease compared to BP + EN, which is mostly due to confusions between CN and EN speakers. From Figure 7, it is noticeable that confusions between languages are rare. Besides that, from a total of 45 wrong classifications, 31% (28) are from BP, 34% (31) from EN, and 45% from CN.

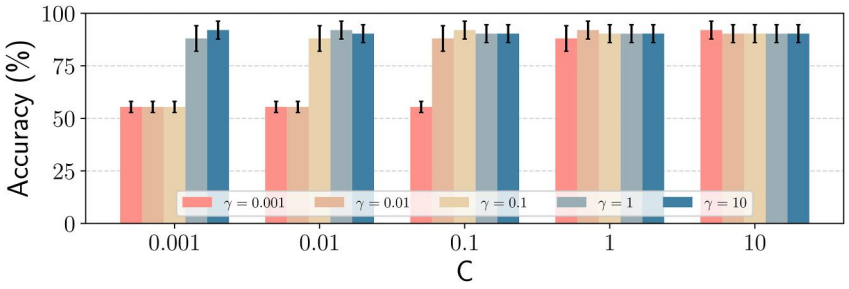


**FIGURE 5** Fine-tuning results for linear SVM using Brazilian Portuguese. SVM, Support Vector Machines

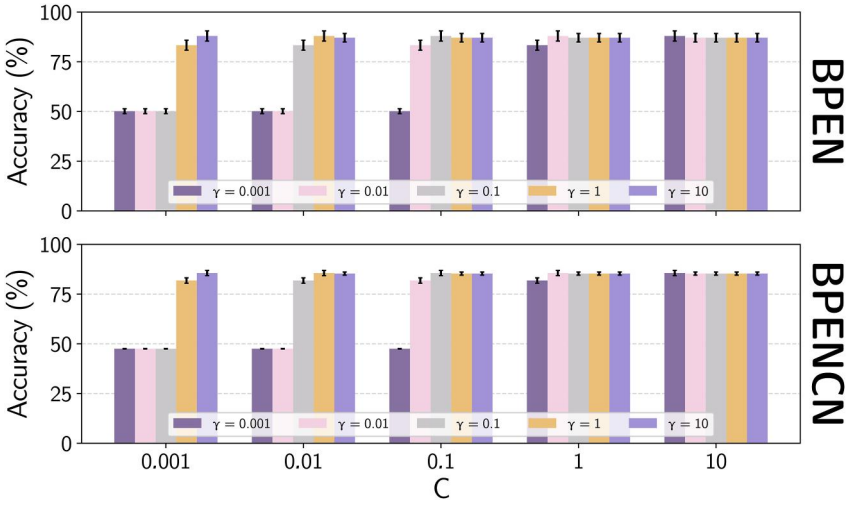


**FIGURE 6** Fine-tuning results for linear SVM for BP + EN (upper) and BP + EN + CN (down). BP, Portuguese; CN, Chinese; EN, English; SVM, Support Vector Machines

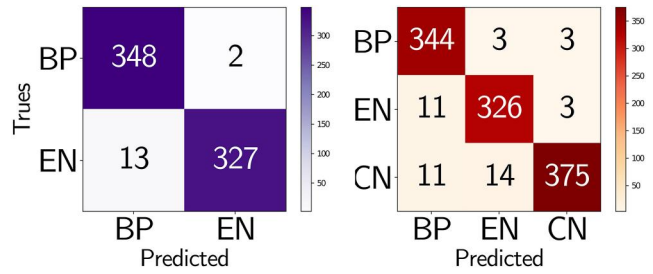


**FIGURE 7** Language×Language confusion matrix for multilingual experiments

### 5.2.1 | Size constraint

Next, a total of 30 tests are evaluated using the same configuration from previous multilingual experiments with smaller versions of the multilingual dataset. As average, these experiments achieved 91.88% ± 1.87%, a 0.12% difference from monolingual results.

### 5.2.2 | Phonemes

At the phonetic level, the most frequent (above the mean of frequencies) phones that appear in both multilingual experiments increase in frequency when Chinese is added (Figure 8 and Figure 9). For instance, 'A', and 'A+' for BP, or 'KCL' and 'IY' for EN. For both languages, there is no phoneme that appears in 2-language that does not appear in 3-language experiments (Figure 10). However, some distinct phonemes occur when using Chinese as additional data in our experiments. These new labels are presented in Table 4.

Brazilian Portuguese has half of most frequent symbols within vowels, followed by consonants. For Chinese, there are 40 sounds (Figure 11). A vowel and two consonants make the top three sounds present in misclassifications, while from 40 phonemes, 21 are vowels.

Meanwhile, the majority of EN phonemes for 2-language setup comes from vowels. These are followed by stops, semi-vowels, and fricatives (Figure 8 right). From these, there are two allophones: 'Q' and 'UX'. With the experiments using three languages, the EN results are consistent with the previous outcomes. The same classes of sounds occur while there is a difference at which ones become more frequent when comparing Figure 8 right and Figure 9 right. Finally, the allophone 'HV' (voiced 'H') is observed when using three languages, as presented in Table 4.

## 6 | DISCUSSION

For better presenting the results, they are segmented into subsections addressing a different aspect of our work.

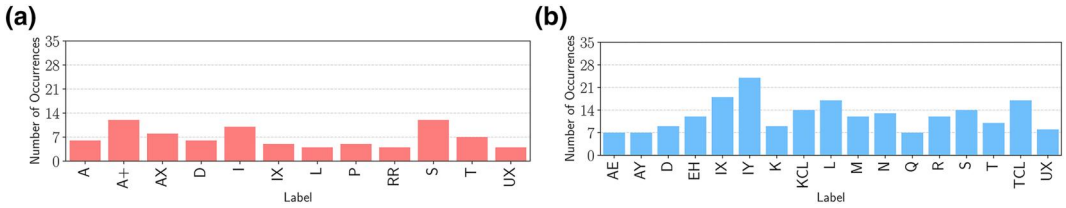


**FIGURE 8** Most frequent phonemes for 2-language experiments on incorrect classifications in BP (left) and EN (right). BP, Portuguese; EN, English

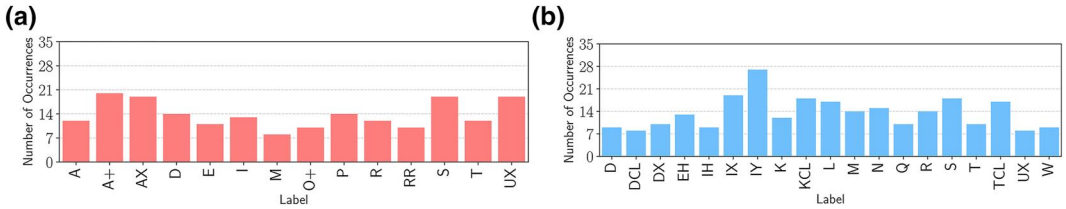


**FIGURE 9** Most frequent phonemes for 3-language experiments on incorrect classifications in BP (left) and EN (right). BP, Portuguese; EN, English

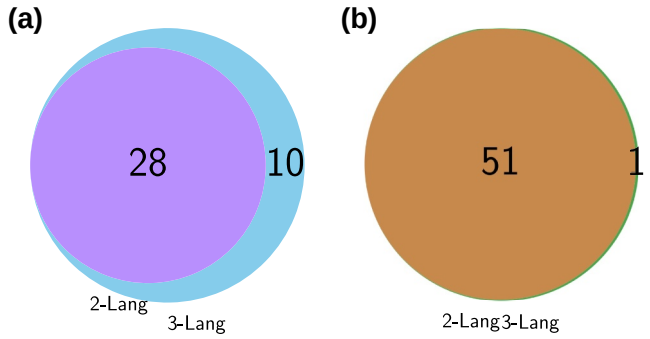


**FIGURE 10** Phonetic intersection between multilingual experimental results in BP (left) and EN (right). BP, Portuguese; EN, English

**TABLE 4** New phonetic occurrences for 3-language experiments

| Language | Labels |
|---|---|
| BP | O~, U+, W~, TJ, SCH, E~, A~+', NJ, U |
| EN | HV |

Abbreviations: BP, Portuguese; EN, English.

## 6.1 | Unsupervised SI

The first model evaluated was the FCM. It was expected to have a bad performance, achieving a maximum of 32.57% using Euclidean distance and $m = 1.5$. As the worst accuracy rate between our classifiers, it is used as our baseline model.

As a more sophisticated version of k-Means, FCM also uses distance measures to separate the data. Its known capacity to recognise well-overlapping data does not help much here. Another element to keep in mind here is the amount of data. Our speakers have a total of 10 samples each, and these consist of short-time recordings [29] and [42]. As a clustering method, this technique does rely on large amounts of data to better separate the feature space. Furthermore, FCM is a non-supervised model, a notable disadvantage when comparing to models that have previous information about classes. This characteristic can have a huge impact on the classification, as unsupervised methods will almost always have worse performance compared to supervise techniques. So, this is another reason for its side-by-side low performance. Possible solutions to improve these results would be to try different feature vectors or even membership functions. The following section discusses the results achieved by our Nearest Neighbours classifiers.

## 6.2 | Nearest neighbours classification

The kNN and FkNN, in contrast to FCM, do not rely on a centre based on all data points. Instead, they manage to create an increasing boundary starting at the unknown sample, rather than matching the new sample to pre-built clusters. Thus, as expected, these models had a much better performance on the experiments when compared to the baseline.

It is true that choosing $k$ too small can lead to an overfit, while making $k \to n$ for a data with $n$ samples would lead to a complete generic model. However, it is important to consider how much data is available; given the small number of samples (6 per speaker) the search space for both crisp and fuzzy versions should be small.

Moreover, the FkNN had a decent score, surpassing its crisp version by 1.42%. Even though not by a large margin, this value can still improve with a few more tuning. Here, several membership functions could be tested to better balance the model according to the data. As pointed out in Ref. [40], different functions can be used, leading to distinct outcomes as shown by their results and thus possibly increasing its accuracy. These further improvements cannot happen for kNN, which had all its parameters changed.

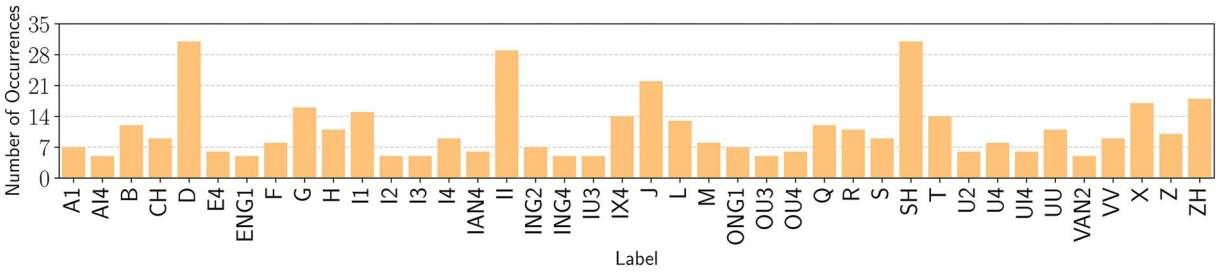


**FIGURE 11** Chinese phoneme occurrences on incorrect classifications

## 6.3 | SVM classification

This classifier managed to obtain the highest accuracy score. It is possible to go further and use the Sigmoid or RBF kernel, which achieved 92.29% recognition rate. As argued in the previous section, we choose the Linear kernel in spite of the higher accuracy of the other kernels.

Another aspect to notice is the Linear kernel overcoming the non-linear kernels in some configurations. The reason behind it is the non-pure Linear kernel. As noted in Section 5, it is simulated with a polynomial kernel with degree 1. While theoretically they are equivalent, their implementations have their divergences. Using this strategy creates a Linear SVM that can be influenced by $\gamma$, which may affect the performance.

Furthermore, the polynomial kernel has a very akin change on accuracy. When comparing Figure 5 with the results in Figure 6 most of them shows a low accuracy for small values of $\gamma$ and $C$. $\gamma$, on the other hand, loses the effect as it and $C$ increase. This accuracy progression is found on both monolingual and multilingual results for this kernel. This can be explained by how sensitive to $\gamma$ each model is. As discussed, the polynomial kernel can be influenced by changes on $\gamma$, which can be visualised in our results (Figure 5). However, its influence decaying shows that this parameter cannot force the borders enough to make individual samples separable from others. Therefore, small values are able to change the border shape by a small factor, while higher values do not make any difference. Although, RBF and Sigmoid are much more sensitive to this parameter, being able to even outline an individual sample inside a distinct class space.

## 6.4 | Identification with fuzzy methods

Fuzzy methods are becoming increasingly popular in several artificial intelligence segments. However, these methods really lack on implementations or even dedicated libraries when compared with their crisp versions. This can restrict the use of these classifiers as the situation creates an environment without standard implementation.

First, the fuzzy classification in this work has two approaches: unsupervised and supervised. In the former, the FCM tries to discriminate the identity of a speaker. This method has a small accuracy rate, but that can still go beyond with additional modifications. For the latter, this work uses a Fuzzy k-Nearest Neighbour, which outperformed its crisp version. Overall, these methods offer a great flexibility. They allow a vast set of characteristics to be modified. Besides, the flexibility on classifying is found better on the performance of FkNN compared to kNN.

Since they require additional computational resources to determine the membership degrees of the current data, they also demonstrate a large disadvantage on performance. It is also true for the FkNN, for which the initialisation used in our work uses a kNN classification. Of course, the SKLearn [41] library has several optimisations for their algorithms, and it becomes quite evident when looking at the train and test duration. Therefore, if performance is crucial in the project, either a very optimised algorithm should be used, or a careful trade-off between performance and accuracy rate should be considered. The next section will discuss the multilingual aspect of this work.

## 6.5 | On the multilingual SI

Results showed that adding a second language reduced the model accuracy by 4.03% and by 6.41% for three, thus indicating that our hypothesis would fail. However, language is not the only variable to consider, due to other characteristics being able to influence SI results. For this problem, results can easily be influenced by gender and number of speakers, the last one being the most harmful for identification problems. As shown in Section 5.2, it is also crucial to consider how the confusions are distributed by language and gender as well as take into account the increasing number of classes as new languages are added.

For confusions, the proportions of speaker languages are fair and do not present any biases. A more careful analysis showed that genders are also balanced for our mistakes, with male speakers appearing on 46% of them and therefore discarding any influence of it on our outcomes.

Furthermore, some information extracted from it is quite interesting. Chinese speakers hold 79.16% of opposite gender mistakes, that is, predicting a male speaker with a female or the other way around. From this proportion, the female speakers represent a large amount. Except by 1 test, every mistake of Chinese speaker by English speaker is between female (CN) and male (EN). For Portuguese, from a total of 9 Chinese females, only two are predicted as female BP speakers. This

suggests that Chinese male voices are very distinct from both BP and EN males. CN females, on the other hand, have a close relation to male voices. But, this conclusion is not a fact, given the presence of some noises on Chinese audio.

Analysing the language misclassifications, they are quite balanced between languages. Taking into consideration the rows and columns in the confusion matrices (Figure 7) the frequency of each language is consistent with their volume of data. There are 340 EN, 350 BP and 400 CN speakers, and the mistakes go as BP, EN, CN. Also, there are not many changes between the confusion matrices from 2 to 3 language experiments. In this transition, EN classifications have an error reduction by 2 for BP column, while BP has increased by 1 at the EN column. Also, note that charts for accuracy rate in Figures 5 and 6 have similar progressions. Besides the influence of $\gamma$ previously discussed, comparing this behaviour among languages may also indicate a similar feature distribution among them.

After concluding that our results are not influenced by gender, experiments were executed to assess the increasing number of classes. Identification problems naturally suffer when there are many classes. This process allowed creating datasets with 34 speakers, one less than the proposed monolingual dataset, by ignoring the gender. This is only possible because gender does not influence our results. Otherwise, this new dataset could be biased. Then, these experiments resulted into 91.88% accuracy. Figure 12 compares our main results from each dataset: monolingual, 2 languages, 3 languages, and 3 languages with reduced number of speakers. This new result is 0.12% lower in comparison with the monolingual accuracy, which is a fairly close score, thus, indicating that consecutive reductions of accuracy in our results are likely due to the increasing number of classes.

At the end, the experiments shift towards an investigation about how each phoneme may affect our results. As shown in Section 5, at all experiments the vowels are the most common, possibly because the languages used here naturally have a high occurrence of vowels in words. Also, every phoneme that appeared in one experiment, is also present on the other. Finally, only one new phoneme is added for English in BP + EN + CN experiments, which is an allophone for voiced 'H', a fricative. For BP, however, there are 10 distinct sound occurrences when Chinese is added. Most of them are vowels, but it is more interesting that except for 'U', the other is closed-mid to open vowels, thus, indicating that while these vowels are very distinct from the EN versions, they have a much closer similarity to Chinese sounds. Therefore, it hardly seems like phonemes are affecting the results. However, it is clear that languages have similar sounds, as adding languages have preserved the phonemes from previous experiments.

Finally, with the arguments and data presented here, the conclusion is that using the methods given in this research study makes speaker identification a language independent task.

## 7 | FINAl REMARKS AND FUTURE WORK

This paper presented results for multilingual closed-set text-independent speaker identification. It is crucial to keep in mind that our objective is not achieving high accuracy. The aim of this paper was to investigate how SI systems behave in multilingual environments. With our results, using the configurations described in this work, speaker identification has certain robustness in a multilingual environment.

As a result of combining different datasets, a lot of our effort was put into preparing and processing the data. Mostly, to make our results more reliable for SI it was crucial to make sure the data would not lead our results towards any language. Also, only the DARPA-TIMIT was made with identity recognition in mind. So, it made necessary to organise both BP and CN to then run our experiments. Of course, all this work on the datasets would be unnecessary if they were consistent with each other or they were a single data. However, no dataset with the required characteristics, inside the budget of this research study, were found. In general, they either have a highly disproportional language distribution or do not include BP. Although, having a decent dataset being found, there would be far more time to experiment different strategies and obtain even more interesting results.

Some segments of this work can be improved or expanded. First, most of our findings come to the conclusion that the model is language independent, but the influence of the feature vector is not investigated. A comparison between different features, such as $x$-vectors or Linear Predictive Coding, could enrich the discussion around multilingual SI. Furthermore, a better method to evaluate the influence of the number of classes could be used. These results are shown in Figure 8 and were obtained through random experiments. A better method would be to split and label each language data, then test all its combinations. This way, one can ensure that all speakers are evaluated.

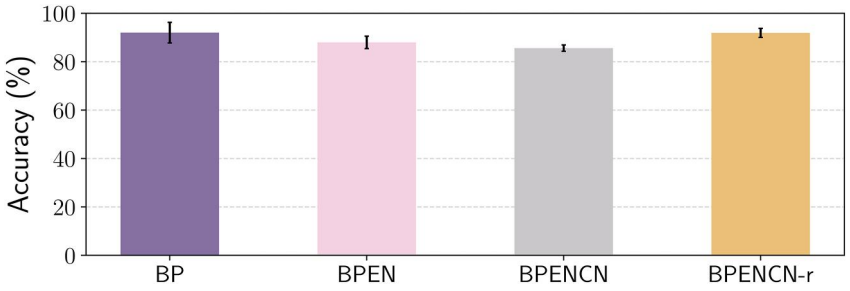


**FIGURE 12** Comparison of monolingual, 2-language, 3-language, and size constraint results

## ACKNOWLEDGEMENTS

This research is sponsored by the Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES) Finance code 001, in which we have no business and/or financial interest.

## CONFLICT OF INTEREST

The authors declare there are no conflicts of interest.

## DATA AVAILABILITY

The data that support the findings of this study are available in MEGA[2]. These data were derived from the following resources available in the public domain: DARPA-TIMIT [43] from Kaggle [33], AISHELL-1 [35] from the OpenSLR[3] repository, and LapsBenchmark16k [34] from FalaBrasil repository.

## ORCID

*Thales Aguiar de Lima* https://orcid.org/0000-0002-1043-8685

*Márjory Cristiany Da-Costa Abreu* https://orcid.org/0000-0001-7461-7570

## REFERENCES

1. Rabiner, L.R., Schafer, R.W.: Theory and applications of digital speech processing, vol. 64. Pearson, Upper, Saddle River (2011)
2. Campbell, J.P.: Speaker recognition: a tutorial. Proc. IEEE. 85(9), 1437–62 (1997). https://doi.org/10.1109/5.628714
3. Lindh, J.: Forensic Comparison of Voices, Speech and SpeakersLinguistics and Theory of Science University of Gothenburg Box 200, SE-40530 Gothenburg. [phdthesis]. Department of Philosophy, University of Gothenburg (2017). http://hdl.handle.net/2077/52188
4. Bonastre, J.F., Wils, F., Meignier, S.: ALIZE, a free toolkit for speaker recognition. In: Proc. of the International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. I–737. IEEE, Philadelphia (2005)
5. Hu, R., Damper, R.I.: Fusion of two classifiers for speaker identification: removing and not removing silence. In: Proc. of the 7th International Conference on Information Fusion, vol. 1, pp. 429–36. IEEE, Philadelphia (2005)
6. Reynolds, D.A.: Speaker identification and verification using Gaussian mixture speaker models. Speech Commun. 17(1), 91–108 (1995). http://www.sciencedirect.com/science/article/pii/016763939500009D
7. Basu, A., et al.: A Novel Minimum Divergence Approach to Robust Speaker Identification (2015)
8. Devika, A.K., Sumithra, M.G., Deepika, A.K.: A fuzzy-GMM classifier for multilingual speaker identification. In: Proc. of the International Conference on Communication and Signal Processing (ICCP), pp. 1514–8. IEEE, Melmaruvathur, India (2014)
9. McLaren, M.L., Mandasari, M.I., van Leeuwen, D.A.: Source normalization for language-independent speaker recognition using i-vectors. In: Proc. of the Odyssey 2012 - The Speaker and Language Recognition Workshop, pp. 55–61. Singapore (2012)
10. Tong, S., Garner, P.N., Bourlard, H.: An investigation of deep neural networks for multilingual speech recognition training and adaptation. In: Proc. of the Interspeech, pp. 714–8. ISCA, Stockholm (2017). http://infoscience.epfl.ch/record/229214
11. Cai, W., Chen, J., Li, M.: Exploring the encoding layer and loss function in end-to-end speaker and language recognition system. In: Proc. of the Odyssey 2018 -The Speaker and Language Recognition Workshop, pp. 74–81. ISCA, Les Sables d'Olonne, France (2018). https://doi.org/10.21437/Odyssey.2018-11
12. Snyder, D., et al.: X-vectors: robust DNN embeddings for speaker recognition. In: Proc. of the International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 5329–33. IEEE, Calgary, AB, Canada (2018)
13. Cheuk, K.W., et al.: Latent space representation for multi-target speaker detection and identification with a sparse dataset using triplet neural networks. In: Proc. Of the Automatic Speech Recognition and Understanding Workshop. ASRU, pp. 358–64. IEEE, Singapore (2019)
14. Anand, A., et al.: Text-independent speaker recognition for ambient intelligence applications by using information set features. In: Proc. of the Inter-national Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications. CIVEMSA, pp. 30–5. IEEE, Annecy, France (2017)
15. Khanum, S., Firos, A.: A novel speaker identification system using feed forward neural networks. In: Proc. of the International Conference on Energy, Communication, Data Analytics and Soft Computing. ICECDS, pp. 3045–7. IEEE, Chennai, India (2017)
16. Shahin, I.: Speaker identification in a shouted talking environment based on novel third-order circular suprasegmental hidden markov models. Circ. Syst. Signal Process. 35(10), 3770–92 (2015). https://doi.org/10.1007/s00034-015-0220-4
17. Ding, I.J., Shi, J.Y.: Kinect microphone array-based speech and speaker recognition for the exhibition control of humanoid robots. Comput. Electr. Eng. 62, 719–29 (2017). https://doi.org/10.1016/j.compeleceng.2015.12.010. http://www.sciencedirect.com/science/article/pii/S0045790615004395
18. Kacur, J.: Modifications of KNN classifier for speaker identification system. In: Proc. of the International Symposium Electronics in Marine. ELMAR, pp. 35–8. IEEE, Zadar, Croatia (2016)
19. Bansal, P., Imam, S.A., Bharti, R.: Speaker recognition using MFCC, shifted MFCC with vector quantization and fuzzy. In: Proational Conference on Soft Computing Techniques and Implementations. ICSCTI, pp. 41–4. IEEE, Faridabad, India (2015)
20. Bansal, P., Imam, S.A.: Performance of speaker recognition system using shifted mfcc, delta spectral cepstral coefficient (DSCC) and Fuzzy techniques. Int. J. Eng. Technol. 7(28), 278–83 (2018). https://doi.org/10.14419/ijet.v7i2.8.10424
21. Rathor, S., Jadon, R.S.: Text independent speaker recognition using wavelet cepstral coefficient and butter worth filter. In: Proc. Of the 8th International Conference on Computing, Communication and Networking Technologies. ICCCNT'08, pp. 1–5. IEEE, Delhi, India (2017)
22. Singh, M., Singh, R., Ross, A.: A comprehensive overview of biometric fusion. Inf. Fusion. 52, 187–205 (2019). https://doi.org/10.1016/j.inffus.2018.12.003. http://www.sciencedirect.com/science/article/pii/S156625351830839X
23. Chung, J.S., Nagrani, A., Zisserman, A.: VoxCeleb2: deep speaker recognition. In: Proc. of the Interspeech, pp. 1086–90. ISCA, Hyderabad, India (2018)
24. Anand, P., et al.: Few Shot Speaker Recognition Using Deep Neural Networks (2019)
25. Li, L., et al.: Cross-lingual speaker verification with deep feature learning. In: Proc. of the Asia-Pacific Signal andInformation Processing Association Annual Summit and Conference. APSIPA, pp. 1040–4. IEEE, Kuala Lumpur, Malaysia (2017)
26. Chen, L., Wu, C.: Crossed-Time Delay Neural Network for Speaker Recognition (2020)
27. Matejka, P., et al.: Analysis of DNN approaches to speaker identification. In: Proc. of the International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 5100–4. IEEE, Shanghai (2016)
28. Przybocki, M.A., Martin, A.F, Le, A.N.: NIST speaker recognition evaluations utilizing the mixer corpora—2004, 2005, 2006. IEEE Trans Audio Speech. 15(7), 1951–9 (2007). https://doi.org/10.1109/tasl.2007.902489
29. Nagaraja, B.G., Jayanna, H.S.: Mono and Cross lingual speaker identification with the constraint of limited data. In: Proc. of the International

[2] https://mega.nz/folder/9lcjySyR#ACXhJ37CDCMsncPUdVH6hA
[3] https://www.openslr.org/33/

Conference on Pattern Recognition, Informatics and Medical Engineering. PRIME, pp. 439–43. IEEE, Salem, Tamilnadu, India (2012)
30. Nagaraja, B.G., Jayanna, H.S.: Efficient window for monolingual and crosslingual speaker identification using MFCC. In: Proc. of the International Conference on Advanced Computing and Communication Systems, pp. 1–4. IEEE, Coimbatore, India (2013)
31. Casanova, E., et al.: Speech2Phone: A Multilingual and Text Independent Speaker Identification Model (2020)
32. de Lima, T.A., Costa-Abreu, M.D.: A survey on automatic speech recognition systems for Portuguese language and its variations. Comput. Speech Lang. 62, 101055 (2020). https://doi.org/10.1016/j.csl.2019.101055. http://www.sciencedirect.com/science/article/pii/S0885230819302992
33. Kaggle: DARPA-TIMIT Speech Dataset (2019). Accessed 17 December 2019. https://www.kaggle.com/mfekadu/darpa-timit-acousticphonetic-continuous-speech
34. FalaBrasil: LapsBenchmark 16k repository. Accessed 17 December 2019. https://gitlab.com/ufpafalabrasil/gitlab-resources#corpora-de-%C3%A1udio-transcrito (2018)
35. Bu, H., et al.: AISHELL-1: an open-source Mandarin speech corpus and a speech recognition baseline. In: Proc. of the 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment. 20, Seoul, South Korea, pp. 1–5, IEEE (2017). https://doi.org/10.1109/icsda.2017.8384449
36. Sadjadi, O.: NIST SRE CTS Superset: A Large-Scale Dataset for Telephony Speaker Recognition. NIST SRE website (2021). https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=933116
37. Jones, K., et al.: Call my Net corpus: a multilingual corpus for evaluation of speaker recognition technology. In: Proc. of the Interspeech. ISCA'17, pp. 2621–4. ISCA, Toronto (2017)
38. Dunn, J.C.: A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J. Cybern. 3(3), 32–57 (1973). https://doi.org/10.1080/01969727308546046
39. Bezdek, J.C., Ehrlich, R., Full, W.: FCM: the fuzzy c-means clustering algorithm. Comput. Geosci. 10(2), 191–203 (1984). http://www.sciencedirect.com/science/article/pii/0098300484900207
40. Keller, J.M., Gray, M.R., Givens, J.A.: A fuzzy K-nearest neighbor algorithm. IEEE Trans Syst. Man. Cybern. 15(4), 580–5 (1985). https://doi.org/10.1109/tsmc.1985.6313426
41. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–30 (2011)
42. Jayanna, H.S., Prasanna, S.R.M.: Variable segmental analysis-based speaker recognition in limited data conditions. Proc. of the International Conference on Image Processing. 2 (2006)
43. Garofolo, J.S., et al.: DARPATIMIT acoustic-phonetic continuos speech corpus CD-ROM. NIST speech disc 1-1.1 (1993)

**How to cite this article:** de Lima, T.A., Da-Costa Abreu, M.C.: Phoneme analysis for multiple languages with fuzzy-based speaker identification. IET Biome. 1–11 (2022). https://doi.org/10.1049/bme2.12078