Improving Prosodic Features Extraction for Tone Detection in Yemba Language

Kenfack Jeuguim Marc Sturm[1] and Paulin Melatagia Yonta[1,2]

[1] Department of Computer Science, University of Yaounde I, Yaounde, Cameroon
{jeuguimmarc,paulinyonta}@gmail.com
[2] IRD, UMMISCO, F-93143, Bondy, France

Abstract. In tonal languages, tone is of vital importance in differentiating between lexical words and grammatical forms. Tone recognition can improve the performance of speech recognition tasks for tonal languages by re-evaluating word hypotheses using tonal information or by including prosodic features in the acoustic model used. Little effort has been made to evaluate the effectiveness of machine learning approaches for tone detection in African low resourced languages. In this paper we propose a selection of prosodic acoustic features to deal with the linguistic specificities of the Yemba language (spoken in Cameroon) for tone detection. The following features, extracted for each frame of a given syllable, are used: pitch, energy, duration, slope of consecutive F0. Experiments have been conducted using multi-speaker models trained with Naive Bayes, LDA, QDA, SVM and decision trees. The decision trees using the cost complexity pruning method gave the best results: an accuracy of 61.82% and a F1 measure of 58.90%.

Keywords: Prosody · Tone Detection · Low Resourced Language · Yemba.

1 Introduction

Prosodic elements, known as prosodemes, play a crucial role in understanding the messages conveyed by speakers. These include tone, accent, rhythm, rate and intonation. In tonal languages, tone is of vital importance in differentiating between lexical words and grammatical forms. For example in Yemba a tonal language spoken in Cameroon, the word "apa" can mean "taro", "bag" or "door", depending on the tonal contour used. In tonal languages, from an orthographic point of view, graphemes called diacritic symbols are defined to indicate the tones, and form an integral part of the language's alphabet. Thus, tone recognition can improve the performance of speech recognition tasks for tonal languages, either by re-evaluating word hypotheses using tonal information from tone detection models, or by directly including prosodic features in the acoustic model used [6]. Many linguistic papers [7] agree that the fundamental frequency is the acoustic correlate of tone par excellence, even though other components may contribute to it. This has thus led to studies using features derived from the acoustic parameters of prosody and machine learning models to detect tones in Mandarin, Thai and Cantonese languages [6,16,13], based on large corpora of prosodically annotated data. However, although most of the languages of sub-Saharan Africa are tonal, little effort has been made to evaluate the effectiveness of these approaches using prosodic features and machine learning on African languages, and to our knowledge, no study of this type has been carried out on Cameroonian languages. In this paper, we focus on the Yemba language. Yemba is a Niger-Congolese language belonging to the Bantu languages and one of the ten Bamileke languages spoken in Cameroon, specifically in the western region and more particularly in the Menoua department. Yemba is used in radio broadcasts on the department's local channels, as well as in the vast majority of local economic activities. Although it was spoken by more than 300,000 people in 1992, making it the country's third largest native language in terms of native speakers [10], few digital resources are available for its study. The writing system used is based on the AGLC (General Alphabet of Cameroonian Languages) [14], from which special characters representing the three tones of the (low tone, high tone, medium tone) language are chosen, as shown in the illustration in Table 1. From a phonological point of view, the unit carrying the tone in the Yemba language is the syllable, and from a graphemic point of view, the tone is marked on a vowel or a nasal consonant in a syllable [2]. In tonal languages, particularly Yemba, the tones of the syllables preceding and following a syllable have an influence on the tone of that syllable. This takes the form of assimilation, tonal propagation and simplification. To detect tones accurately, a robust tone detection model must be able to distinguish tones in groups of words that differ orthographically only in the diacritical symbols of the tone. A challenge with these languages is that most of these words are either bisyllabic or monosyllabic. For example, the word 'apa' has two syllables, but the preceding syllable, necessary for the tonal context of the syllable 'a', is absent. Hence the need to propose a solution to overcome this challenge. In this paper, we propose to evaluate and compare the performance of a multi-speaker model trained using generative classification algorithms (Naive Bayes, LDA, QDA) and discriminative classification algorithms (SVM and decision trees) on the YembaTone corpus [3] annotated at the tonal and syllabic level of the Yemba language. We use prosodic acoustic features from [6], adapting the feature extractor to take into account certain specificities of the Yemba.

The rest of this paper is organised as follows: Section 2 discusses the related work of other researchers. In Section 3, we details our methodology. In Section 4 we presente our experiments and the results obtained. Section 5 concludes the paper.

2 Related work

Prosodemes are realised by involving the intensity, quantity, duration and pitch of the sound. In [13] Satravaha performs a syllable-segmented classification of tones in Thai speech (a tonal and monosyllabic language) that incorporates the effects of tonal coarticulation, accentuation and intonation using a multilayer

[3] YembaTone is available here: `http://dx.doi.org/10.17632/cx268tmrwn.1`

Table 1: Illustration and impact of the 3 tones of the Yemba language on the second syllable of the apa grapheme sequence. A simple change in the tone of the second syllable implies a direct change in the word

Word in Yemba	**Apa**	**Apā**	**Apá**
Tone of Each Syllable	low-low	low-medium	low-high
Word in English	The bag	The taro	The door leaf

perceptron (MLP) with derived prosodic acoustic features as the input vector. Satravaha showed that they are fairly representative of each of these three prosodemes on tone in the Thai language. He constructed a corpus from 5 male and 3 female speakers. The training set consisted of 100 sentences per speaker comprising 4 monosyllabic words with different stress and tone patterns. The test set consisted of 115 sentences with the same structure as those in the training set. Using ANOVA tests, Satravaha showed that duration and normalized energy are distinctive features between stressed and unstressed syllables and that F0 is not. Based on the hypothesis that intonation in the Thai language is strongly characterised by the gradual decay of the F0 contour, he showed, that the mean F0 in each tone can be used to take into account for the effect of intonation. Based on the hypothesis that tone is characterised by the F0 contour normalised at the level of a syllable, Satravaha found that the realisation of a tone at the level of a syllable can be affected by the realisation of the tone of the following syllable. Moreover it can also affect the realisation of the tone of the preceding syllable. From these analyses, a sequence of features derived from the F0 in which the normalised F0 contour of the current syllable, the previous syllable and the next syllable to take into account the effects of tonal coarticulation were used in the training vector. Then the mean normalised F0 of the syllable and its order number in the syllable to take into account the effect of intonation were added to this input vector. Finally, the degree of syllable stress, to take into account the fact that the normalised F0 contour of stressed and unstressed syllables are different, was added to the input vector, reducing the size of the final vector to 48 . For each speaker, a MLP classifier was trained on each of the 100 training sentences and tested on the 115 test sentences, obtaining an average accuracy rate of 91.36% for the 8 speakers. However, the human explicability and interpretability of the model's decision based on these features poses a challenge, particularly when artificial values of -1 are used for the preceding and following syllables at the beginning and end of the sentence.

Another relevant study conducted in [16] adopted, different normalization schemes on these prosodic features almost similar to the previous study to make them robust to conditions variations on these features that can be modeled as an affine function by exploiting the concept of affine invariance [11]. Indeed, as in [13], Qiao et al emphasise that the realisation of the prosodic model of the different tones varies significantly according to the utterances produced under different conditions. This can be influenced by the speaker, the speaker's gender, the speech rate, the speech style, as well as the speaker's emotion and prosodic state. Prosodic state is used to characterise the prosodic behaviour of a syllable and is related to factors such as the variation in intonation caused by the position of the syllable in the utterance and the coarticulation effect of preceding and following syllables. Qiao et al seek to make the acoustic characteristics of prosody invariant to changes in conditions, which can be modelled by affine transformations. First, they proposed a z-score normalisation of the pitch of the current syllable relative to the syllables in its utterance. Qiao et al showed that if transformations in pitch range and pitch level along an utterance between two speakers can be modelled by affine transformations, then any feature function applied to this normalisation automatically becomes an affine invariant. For pitch conditions that vary only at the syllable level, such as speaker emotion and context intonation, they showed that if these changes can be modeled by an affine transformation, then normalizing the pitch by subtracting the syllable-level mean and dividing by the utterance-level standard deviation makes the pitch robust to variations in syllable-level pitch (but not to variations in syllable-level pitch range) . As far as duration-related features are concerned, they shown that defining duration-related features as the ratio between adjacent syllables makes them invariant to transformations of speech rate conditions (sound) which can be modelled by an affine transformation. Finally, for energy, the logarithmic difference in energy between adjacent syllables is a feature that is invariant to change in loudness, which can be modelled as an added bias on the logarithm of the energy of the syllables in an utterance. This is how they constructed a vector of 21 of these prosodic features, with which they trained an architecture using two linear kernel SVMs. The first is trained for a five-class multiclass classification and provides the a posteriori vector of the syllable, which they call the posteriogram. The second takes as input the posteriograms of the current, previous and next syllable, which it concatenates with the 21 initial prosodic features to be trained to finally predict the tone of the current syllable, allowing a classification of multi-speaker tones at syllable level in Mandarin. They used the large COSPRO-01 corpus1 [15] of Mandarin multi-speaker speech, produced by 38 male and 40 female native speakers, with over 60,000 syllables for training, as well as COSPRO-02, produced by two male and two female native speakers, with just over 10,000 syllables. Both corpora were prosodically annotated and syllable segmented manually, and achieved 68% accuracy on the test dataset. However, the interpretability and explicability of the model's decision is also problematic here, as the authors replace the posteriogram of the preceding and following syllables of the beginning and ending syllables with an a posteriori vector uniformly sampled from a uniformly distributed distribution over the five classes.

A previous study in [6] considerably extends the prosodic feature vector of the previous study, while retaining a large part of the previous vector. Thus, the vector increases from a size of 21 to a size of 52. This extension enables them to compare the performance of a maxout neural network and an SVM for recognising tones in continuous Mandarin speech, using the syllable as the unit carrying the tone. Indeed, when extracting features, Chen et al start from the observation of the phonological properties of the Mandarin language, where most words are monosyllabic or dissyllabic. For each syllable, it is necessary to take into account the influence of its context. The preceding syllable has more influence on the tone of a syllable than the following syllable. This is why they choose the two preceding syllables and the following syllable as the contextual window for a given syllable. For each syllable, they extract 16 pitch-related features. Chen et al begin by dividing the logarithm of the syllable pitch into three segments of equal length, and then extract the mean and slope of the linear approximation of the pitch contour of these three segments. To take context into account, they also extract the mean and slope of the linear approximation of the last segment of the previous syllable and the first segment of the next syllable. They also include the pitch value of the first and last frames, the minimum and maximum values, and the value of the first and last frames adjacent to the pitch of the current syllable. For the four duration-related features, they extracted the duration of the current syllable, its normalised duration (z-score), and the two duration ratios with adjacent syllables. For the six energy-related features, they extract the minimum, mean, maximum, range, standard deviation and root mean square of the logarithm of the energy for the current syllable. For each of these features, Chen et al finally construct dynamic features of the pop-up window. In their experiments, they used the large ASCD corpus, a Mandarin corpus manually labelled for the prosody of continuous speech, spoken by five male and five female speakers. The corpus contains a total of 79,679 syllables, unbalanced across the five tone classes. For training, validation and testing, they used 50,237, 9,820 and 16,628 syllables respectively, selecting a specific proportion of each speaker. They then trained multi-speaker tone detection models, achieving an accuracy of 78.21% with the maxout network and 74.19% with the SVMs. However, it should be noted that one of the limitations of their work lies in the fact that the samples of utterances in the training, validation and test datasets are not mutually exclusive in terms of speakers. This situation could lead to unrealistic performance estimates for a model that is supposed to adapt to new speakers.

3 Methodology

In this section, we present the different steps for evaluating and comparing the performance of the combination of prosodic acoustic features from the literature and classical machine learning classification algorithms, with a view to solving the task of tone detection in the Yemba language. These steps range from the acquisition of the dataset to the choice of performance evaluation metrics. Figure 1 shows the general architecture of the system. The figure is described in the next sections.

3.1 Dataset

We developed a pronunciation dataset of specific isolated words in the Yemba language. These words were segmented into syllables and prosodically labelled according to tone, using Praat software [4]. This dataset contains a total of 6754

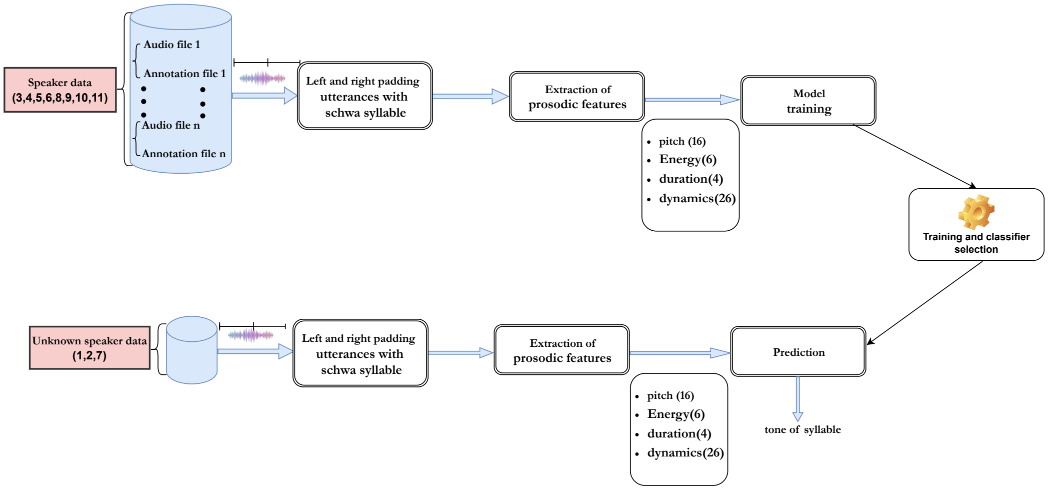

Speaker data (3,4,5,6,8,9,10,11)
Audio file 1
Annotation file 1
Audio file n
Annotation file n
Left and right padding utterances with schwa syllable
Extraction of prosodic features
Model training
pitch (16)
Energy(6)
duration(4)
dynamics(26)
Training and classifier selection
Unknown speaker data (1,2,7)
Left and right padding utterances with schwa syllable
Extraction of prosodic features
Prediction
pitch (16)
Energy(6)
duration(4)
dynamics(26)
tone of syllable

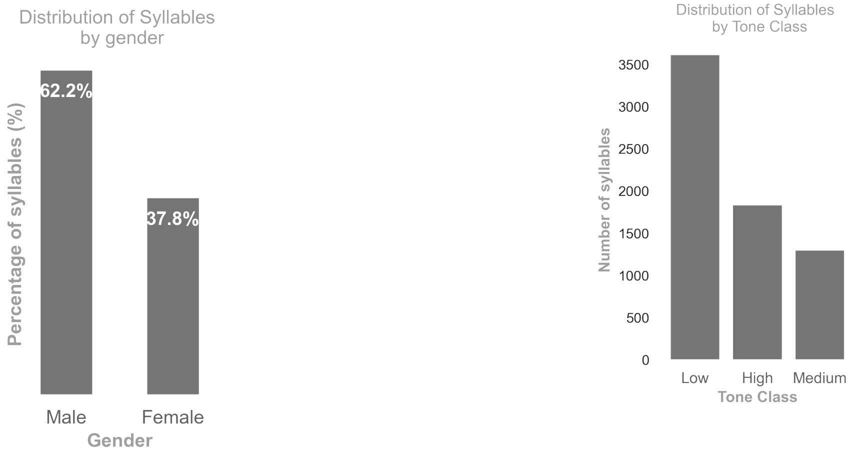
Fig. 1: Experiment Setup

pronunciations of syllables. Its particularity lies in the fact that it results from the pronunciation of 11 native speakers, including 4 men and 7 women, aged between 13 and 50. Most of them are masters-level or higher language students, which means they have mastered the sounds in a dictionary of isolated words that we have compiled in collaboration with linguists. This dictionary contains groups of words whose only difference is the diacritical symbol for tone. This differentiation was designed to create sufficiently discriminative tone detection models. An in-depth descriptive analysis of the dataset is presented in Figure 2a and Figure 2b. For more details on the data collection protocol, please consult the online repository. It should be noted that the sampling frequency used was 44.1 kHz, with a resolution of 16 bits.

3.2 Feature extraction

Our prosodic acoustic feature extraction module is inspired, as mentioned earlier, by the work of [6]. In this section, we present these features and the choices that have been made to adapt their approach to our own data.

Syllabic context adaptation In this paper, we are working with a corpus of isolated words, so there is no information about their natural context. To fix the issue of missing context for syllables at the beginning and the end of utterance, we have borrowed the schwa syllable from English. Ended the neutral syllable has not been identified in the Yemba language. English is characterised by its tonic accent, where certain syllables within words are pronounced with more stress than others. In English, syllables within words can be classified into three levels of stress: stressed, secondarily stressed or unstressed. The syllable



(a) (b)

Fig. 2: On the left, the distribution of syllables according to gender, and on the right, according to tone. Given the limited size of the data and the imbalances, we have made some crucial choices, including the cross-validation strategy and the evaluation metrics, to ensure both a reliable assessment of performance and the robustness of the model.

schwa, predominant in the English language, is often considered to be neutral or weak in stress and tone. It has no tone of its own and exerts no tonal effect on surrounding syllables. This makes it a suitable option for use as a substitute in the missing context of certain syllables in Yemba, as it is unlikely to influence the tone of these syllables. This is why each speaker has articulated the syllable schwa, and we use its enunciation as padding at the beginning and end of words.

Acoustic characteristics related to pitch. For each syllable, we obtain a sequence of fundamental frequencies by analysing overlapping frames. The algorithm used to estimate the fundamental frequency in a frame is based on the method proposed by [3] and is implemented in the Praat software. The main idea behind this approach is to choose the candidate for the fundamental frequency from the local maxima of the frame autocorrelation estimate.

As proposed in [6], pitch-related features are extracted as follows: for each syllable, 16 features are extracted. First, the log-F0 sequence of the current syllable is segmented into three segments. The slopes of the linear approximation of the 3 segments of the current syllable are extracted. In addition, the slope of the linear approximation of the last segment of the previous syllable and that of the first segment of the next syllable are extracted. The average of the linear approximation of the 3 segments of the current syllable is also extracted, as is the average of the linear approximation of the last segment of the previous syllable and the first segment of the next syllable. The maximum and minimum pitch values of the current syllable are extracted, as are the pitch values of the last voiced frame of the current syllable and the first voiced frame of the current syllable. In addition, the pitch values of the last voiced frame of the previous syllable and the first voiced frame of the next syllable are also extracted.

Here, the Mean of an affine function f on an interval $[a, b]$ is calculated as follows:

$$\mu_{[a,b]}(f) = \int_a^b f(x)dx = \frac{f(a) + f(b)}{2} \tag{1}$$

Acoustic characteristics linked to energy. Similar to the pitch characteristics, the energy is also extracted for each frame of a given syllable. In the context of a signal x in a frame running from time sample t_1 to sample t_2, the energy is defined as follows:

$$Energy = \sum_{t=t_1}^{t_2} x^2[t] \tag{2}$$

As proposed in [6], energy-related features are extracted as follows: for each syllable, 6 features are extracted. The maximum value, minimum value, range, mean, root mean square and standard deviation of the energy of the current syllable are extracted, all with the energy scaled logarithmically.

Characteristics relating to duration. As proposed in [6], duration-related features are extracted as follows: 4 duration-related features are extracted. The duration of the current syllable (in seconds), the normalised duration of the current syllable and the duration ratio with respect to the following and current syllables are extracted.

Calculation of dynamic characteristics As proposed in [6], they are normalisations of the 26 individual characteristics of the current syllable with respect to the equivalent statistics in the wider context. In our implementations, we choose the utterance, more precisely the pronunciation of the isolated word, as the contextual window.

3.3 Classification algorithms

In the following section, we will briefly present three generative classification algorithms and two classical discriminative methods. These approaches are frequently used to solve classification problems and will be discussed in terms of their relevance and effectiveness.

Generative Classification algorithms: Naïve Bayes, Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) are algorithms that assume that class likelihoods follow a multivariate Gaussian distribution, while priors follow a multinomial distribution. These methods often prove to be a wise choice for initiating a classification task, thanks to their rapid execution and relative simplicity. QDA, for example, assumes that class likelihood densities are characterised by individual covariance matrices, leading to quadratic decision functions. The LDA approach, on the other hand, is based on the assumption of a common covariance matrix for the likelihood distributions of all the classes, leading to linear decision functions. In contrast, Naïve Bayes follows a similar logic to QDA, but makes the assumption that features are independent, leading to diagonal covariance matrices. In practice, it should be noted that QDA can have limitations when the dimensionality of the examples is high and the number of samples per class is small. To obtain robust estimates with small sample sizes, one solution may involve aggregating the data and estimating a common covariance matrix, an approach adopted by LDA. Another strategy is to make specific assumptions about the features, as is the case with Naïve Bayes.

Discriminative Classification Algorithm: Support Vector Machines (SVM) is a machine learning algorithm that aims to find the best hyperplane that maximises the margin between two classes in the input space. The optimisation problem to be solved is :

$$\begin{aligned} & \underset{w,b,\xi}{min} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \\ & \text{sc } y_{(i)}(w^T.x^i + b \geq 1 - \xi_i), i = 1, \cdots, N \\ & \ \xi_i \geq 0 \end{aligned} \tag{3}$$

The parameters w and b represent the separation hyperplane. The soft variables ξ_i are used to quantify how the noise in the training data is taken into account, in order to prevent the margin from over-adapting to disturbances in the training data. This over-adaptation could lead to over-fitting. C is the regularisation parameter. It defines the compromise between maximising the margin and minimising the classification error. For a large value of C, the algorithm strives to reduce the error in the learning phase. For a small value of C, the model penalises the theta parameters, which may lead to a very simple model that may not be suitable. Kernel functions are one of the major tricks of the SVM trade. Proposed by Vapnik, these functions are used when the data cannot be separated linearly. They are generally interpreted as measures of similarity between samples from the point of view of the application.

A decision trees is a non-parametric supervised learning algorithm. It has a hierarchical tree structure, consisting of a root node, branches, internal decision nodes and terminal leaves. The growth process of a decision tree can be expressed as a recursive algorithm with following steps: Choose the feature so that when the parent node is split, the result is greater information; Stop if the child nodes are pure or if it is no longer possible to improve the purity of the class; Return to the first for each of the two child nodes. Information gain is the objective function to be maximised at each division and is defined as follows:

$$IG(D_p, f) = I(D_p, f) - \sum_{j=1}^{p}\frac{N_j}{N_p}I(D_j) \tag{4}$$

f is the feature on which the division is performed, D_j the child nodes, D_p the parent node. I is the impurity measure. N_p is the number of training examples in the parent node, N_j the number of examples in the j^{th} child node. The literature distinguishes three measures of impurity: entropy, Gini index and classification error. However, in the presence of noise in the data, by letting the tree grow until it is at its purest, we can obtain very large trees that overfit the training data. To avoid this, we can use pruning methods. On the one hand, pre-pruning: maximum depth, minimum number of samples per node and minimum split improvement. Secondly, post-pruning: cost complexity pruning more details about these pruning can be found in [5].

3.4 Performance evaluation metrics

The dataset shows a significant imbalance, with the "low tone" class largely predominating. Consequently, our evaluation metric when selecting the model must take account of this asymmetry. The model must be able to detect all the correct samples in a given class (recall), while classifying only the correct samples in that class (precision). This is why we opted for the F1 measure, which is recommended for balancing the trade-offs between optimising precision and recall. More specifically, given that we are working with a multi-class classification model, we will adopt the weighted macro-mean method for aggregation. This method takes into account class imbalance, unlike other aggregation approaches in the context of multiclass classification.

$$F1_{\text{macro}} = \frac{\sum_{i=1}^{C} w_i \cdot \text{F1}_i}{\sum_{i=1}^{C} w_i} \tag{5}$$

C represents the number of classes, w_i is the weight associated with class i and $F1_i$ is the $F1$ measure for class i.

4 Experiments

4.1 Experimental protocol

In our experiments, the YembaTones audio files were divided into two separate sets: a training set (including 8 speakers) and a validation set, and a test set (involving 3 speakers). These sets were distributed according to the configuration shown in the table 2.

To estimate the acoustic properties within a frame, we apply the Hamming window with a duration of 25 ms and an overlap of 20 ms between successive frames. In addition, in order to access the annotation information and to estimate certain acoustic characteristics, in particular the fundamental frequency F0, which come from Praat, we use the Parselmouth [9] Python library.

For each classifier, we select the best hyperparameters using a factorial design with scikit-learn's GridSearch tool. The cross-validation strategy chosen for model selection is the 5-fold stratified group. We first use a 5-fold division due to the limited size of our dataset. Second, we apply stratification due to the imbalance of the samples with respect to the target attribute. Finally, we adopt a

Table 2: Distribution of the Number of Syllables in the Training and Test Sets. In our experiments, the training data includes samples from speakers 3, 4, 5, 6, 8, 9, 10,11, while the test set includes samples from speakers 1, 2, and 7.

	Training Set			**Test Set**		
Tone	**Low**	**High**	**Medium**	**Low**	**High**	**Medium**
Number of Syllables	2760	1380	990	856	458	310

cluster approach to allow our model to demonstrate the best performance while measuring its resilience to speaker changes. Each group corresponds here to a speaker.

The first experiments consist of using models assuming a Gaussian distribution of the data. The controlled factor is the selection, in the grid, of one of the following algorithms: LDA, QDA or Naïve Bayes.

The second experiments consist of choosing the SVM algorithm as a starting point, then taking into account its various hyperparameters as controllable factors: the regularisation parameter C, the kernel and parameters. Following the recommendations in the literature[16,1,8], the different search grids we used are detailed in the Table3. For experiments with decision trees, given that the Gini impurity measure and entropy generally produce similar results, and that the misclassification error does not contribute significantly to tree growth due to its low sensitivity to variations in probabilities [12], we opt to use Gini impurity in our experiments. Given that the choice of hyperparameters for pruning with the other methods is almost intuitive, we only consider post-pruning with the ccp [5] method, whose optimal subtrees in terms of maintaining the compromise between minimising classification error and the number of leaves are clearly defined in the literature. Our approach mainly consists in obtaining the effective values of on the training dataset, and then selecting the parameter as well as our final model by cross-validation on this list of values. This approach allows us to strike the right balance between training and model generalization.

Table 3: Hyperparameter Search Grids for SVM Experiments

Grid Search	C Range	Linear Kernel	RBF Kernel	RBF Gamma	Feature Normalization (min-max)	References
1	$\{10^{-3}, \ldots, 10^{3}\}$	Yes	Yes	$\{10^{-3}, \ldots, 10^{3}\}$	No	[16],[1] for C
2	$\{2^{-5}, \ldots, 2^{15}\}$	Yes	Yes	$\{2^{-15}, \ldots, 2^{3}\}$	No	[8] for C and γ
3	$\{2^{-5}, \ldots, 2^{15}\}$	Yes	Yes	$\{2^{-15}, \ldots, 2^{3}\}$	Yes	[8] for C and γ
4	$\{10^{-3}, \ldots, 10^{3}\}$	Yes	Yes	$\{10^{-3}, \ldots, 10^{3}\}$	Yes	[16],[1] for C

The model learning experiments were carried out using the scikit-learn machine learning library. In the interests of reproducibility, the random number generation seed for all the libraries was set at 12321. The experiments were carried out on a Colab server with a 2.2 GHz CPU with 2 cores and 12 GB of RAM.

4.2 Results and discussions

For our various experiments, the performance and parameters of the best models obtained after cross-validation for each configuration, as described in the protocol, are shown in the Table 4.

Table 4: Results of different classifiers with their best configurations.

Classifier Type + Grid	**Best Model and Parameters**	**F1 on Test Set (%)**	**Accuracy (%)**
Generative Classifiers	LDA	**50.18**	**55.42**
	Naïve Bayes	46.97	54.00
	QDA	44.35	53.63
SVM + Grid 1	{’C’: 1.0, ’gamma’: 0.01, ’kernel’: ’rbf’}	**54.83**	**55.60**
SVM + Grid 2	{’C’: 2.0, ’gamma’: 0.015625, ’kernel’: ’rbf’}	51.34	55.42
SVM + Grid 3	{’C’: 1.0, ’gamma’: 1.0, ’kernel’: ’rbf’}	53.99	55.48
SVM + Grid 4	{’C’: 1.0, ’gamma’: 1.0, ’kernel’: ’rbf’}	53.99	55.48
Decision Tree + Grid provided by the effective α values of CPP	{’ccp_alpha’: 0.0029}	**58.87**	**61.82**

The results presented highlight significant distinctions between algorithms based on Gaussian distribution assumptions for data modelling, namely Naive Bayes, LDA and QDA. More specifically, the LDA algorithm stands out by displaying superior performance, reflected by an average F1 score of 50.12% and an accuracy of 55.42%. Further analysis of these results suggests a potential correlation with the limited size of the training dataset. The limited size of our dataset may have restricted the ability of the algorithms to perform accurate estimates of individual covariance matrices, particularly in the context of QDA, which requires such class-specific estimates. From this perspective, the QDA algorithm could have been disadvantaged by a lack of data conducive to the generation of reliable estimates, which would have potentially impacted its performance. In contrast, the LDA algorithm capitalises on the use of a shared covariance matrix, giving it an advantage in terms of covariance estimation, thanks to the availability of a relatively larger quantity of data. However, our results also raise the question of the inductive bias introduced by the Gaussian assumption on the distribution of the data. This observation is reinforced by the performance of SVMs, which outperforms that of LDA independently of the search grid. The optimal SVM model is obtained with C=1, an "rbf" kernel with parameter gamma = 0.01, and without feature scaling, resulting in an average F1 score of 54.83% and an accuracy of 55.42%. The SVM results on different grids show that linear kernels failed to separate the data more effectively using prosodic features. Furthermore, the lack of improvement due to scaling suggests that each prosodic feature variable does not contribute equally to the tone classification process in the Yemba language. This could indicate that some pitch-related features are of greater importance, while others, such as duration and energy, act as accompanying factors to account for contextual variations, such as the speaker's emotional state or speech rate. This hypothesis is further supported by the fact that the decision tree-based model performs significantly better than the other algorithms, with an average F1 score of 58.87% and an accuracy of 61.82%. This finding can be attributed to the fact that decision trees do not assume complex assumptions and can act as feature selectors, reinforcing the idea that extracted prosodic features, without complex combinations as in SVM models, offer a better ability to discriminate tones. The improved performance of decision trees also suggests the possibility of exploring experiments with ensemble models with decision trees as the base learner, with the potential aim of further improving tone detection using prosodic features.

5 Conclusion

In this study, we explored a crucial aspect of speech recognition in tonal languages, namely the detection of tones, which play a fundamental role in distinguishing between lexical words and grammatical forms. Previous work on tonal languages, such as Mandarin or Thai, could not be directly applied to our study. Our focus was on Yémba, a tonal language spoken in Cameroon. We developed a comprehensive approach to tone detection in Yemba using various machine learning algorithms, including Naive Bayes, LDA, QDA, SVM and decision trees. Faced with the challenges posed by the linguistic properties of Yemba, we carefully selected prosodic acoustic features such as pitch, energy, duration, slope and the average of consecutive F0 values. By filling in the left and right of the utterances with the neutral syllable "schwa" from English, we solved the problem of missing syllables at the beginning and end of the utterance. These features were applied to a YembaTone corpus that we specially created and published for this study. Our experiments showed that the decision trees, using the cost-complexity pruning method, gave the most promising results, with an accuracy of 61.82% and an F1 measure of 58.9%. Ultimately, our study highlights the importance of taking linguistic peculiarities into account when designing and evaluating machine learning models for tone detection. Moreover, the remarkable performance of the decision trees suggests that these prosodic features, even without complex combinations, can be used to distinguish tones in the Yemba language. This paves the way for future experiments with ensemble models using decision trees as a base learners, such as rendom forests, in the hope of improving performance. Furthermore, it would be judicious to integrate these prosodic features to improve speech recognition in the Yemba language.

6 Acknowledgement

This work has been funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 101007666, the Agency is not responsible for these results or use that may be made of the information.

References

1. Alpaydin, E.: Introduction to Machine Learning, pp. 349–385. Adaptive Computation and Machine Learning, MIT Press, Cambridge, MA, 3 edn. (2014)
2. Barreteau, D., Hedinger, R.: Description des langues camerounaises, pp. 239–269. ORSTOM ; ACCT (1989)
3. Boersma, P.: Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. Proc. Inst. Phonetic Sci. Univ. Amsterdam **17**, 97–110 (1993)
4. Boersma, P.: The Use of Praat in Corpus Research. In: The Oxford Handbook of Corpus Phonology, pp. 342–360. Oxford University Press (05 2014)
5. Breiman, L., Friedman, J., Stone, C., Olshen, R.: Classification and Regression Trees. Taylor & Francis (1984)
6. Chen, M., Yang, Z., Liu, W.: Deep neural networks for mandarin tone recognition. In: 2014 International Joint Conference on Neural Networks (IJCNN). pp. 1154–1158 (2014)
7. Compaore, L.: Essai d'analyse de la prosodie du Mooré : ton et intonation. Theses, Université Sorbonne Paris Cité (Jul 2017)
8. Hsu, C.W., Chang, C.C., Lin, C.J.: A practical guide to support vector classification. Tech. rep., Department of Computer Science, National Taiwan University (2003)
9. Jadoul, Y., Thompson, B., de Boer, B.: Introducing parselmouth: A python interface to praat. Journal of Phonetics **71**, 1–15 (2018)
10. Kouesso, J.R.: The yemba language (cameroon): 90 years of tone orthography. International Journal of African Society, Cultures and Traditions **4**(3), 1–15 (August 2016)
11. Qiao, Y., Suzuki, M., Minematsu, N.: Affine invariant features and their application to speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2009, 19-24 April 2009, Taipei, Taiwan. pp. 4629–4632. IEEE (2009)
12. Raschka, S., Liu, Y., Mirjalili, V., Dzhulgakov, D.: Machine Learning with PyTorch and Scikit-Learn: Develop Machine Learning and Deep Learning Models with Python. Expert insight, Packt Publishing (2022)
13. Satravaha, N.: Tone classification of syllable-segmented thai speech based on multilayer perceptron. Tech. Rep. 1611, Graduate Theses, Dissertations, and Problem Reports (2002)

14. Tadadjeu, M., Sadembouo, E.: General Alphabet of Cameroon Languages. Université de Yaoundé, SIL Internationale (1979)
15. Tseng, C.y., Cheng, Y.c., Chang, C.H.: Sinica cospro and toolkit—corpora and platform of mandarin chinese fluent speech. In: Oriental COCOSDA 2005 (December 6-8 2005)
16. Wang, Y.B., Lee, L.S.: Mandarin tone recognition using affine-invariant prosodic features and tone posteriorgram. In: Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech 2010). pp. 2850–2853 (September 2010)