Standard Yorùbá Context Dependent Tone Identification Using Multi-Class Support Vector Machine (MSVM)

*SOSIMI, AA.; ADEGBOLA, T.; FAKINLEDE, OA.

Department of Systems Engineering, University of Lagos, Àkọkù, Lagos, Nigeria
*Corresponding Author Email: asosimi@unilag.edu.ng, taintransit@hotmail.com, oafak@unilag.edu.ng

ABSTRACT: Most state-of-the-art large vocabulary continuous speech recognition systems employ context dependent (CD) phone units, however, the CD phone units are not efficient in capturing long-term spectral dependencies of tone in most tone languages. The Standard Yorùbá (SY) is a language composed of syllable with tones and requires different method for the acoustic modeling. In this paper, a context dependent tone acoustic model was developed. Tone unit is assumed as syllables, amplitude magnified difference function (AMDF) was used to derive the utterance wide $F_0$ contour, followed by automatic syllabification and tri-syllable forced alignment with speech phonetization alignment and syllabification SPPAS tool. For classification of the context dependent (CD) tone, slope and intercept of $F_0$ values were extracted from each segmented unit. Supervised clustering scheme was utilized to partition CD tri-tone based on category and normalized based on some statistics to derive the acoustic feature vectors. Multi-class support vector machine (MSVM) was used for tri-tone training. From the experimental results, it was observed that the word recognition accuracy obtained from the MSVM tri-tone system based on dynamic programming tone embedded features was comparable with phone features. A best parameter tuning was obtained for 10-fold cross validation and overall accuracy was 97.5678%. In term of word error rate (WER), the MSVM CD tri-tone system outperforms the hidden Markov model tri-phone system with WER of 44.47%.

DOI: https://dx.doi.org/10.4314/jasem.v23i5.20

Copyright: Copyright © 2019 Sosimi et al. This is an open access article distributed under the Creative Commons Attribution License (CCL), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Dates: Received: 21 December 2019; Revised: 20 May 2019; Accepted 25 May 2019

Keywords: Syllabification, Standard Yorùbá, Context Dependent Tone, Tri-tone Recognition

In recent times Automatic Speech Recognition (ASR) has been of special interest to researchers; its application domain has also expanded from simplest system of digit recognition to portable cross-language spontaneous dialogue systems, such development is mainly due to the improvement in computational power and modeling approaches for representing speech signal. While significant progress have been accomplished in phone language ASR, there are still large number of issues that have not been solved, particularly for under-resource languages, where annotated speech resources are limited (Eme and Uba, 2016). Tone languages denote a large proportion of the spoken languages of the world and yet lexical tone is an understudied features. This is attributed to the unsettled questions on building of the vocabulary, what should constitute the sub-word units, how structures over these units are parameterized, modeled and trained. In languages such as SY, tone forms an integral element of the syllable and serves an essential function in distinguishing meaning of syllables with same phonological configuration. Tonal languages have distinctive tones and the number of tones differs across languages. For example, SY, Thai, Cantonese, and Hausa have three, five, nine and two lexical tones respectively. Hence, tone languages, such as Standard Yorùbá, differ from other tone languages, for instance, in some Asian languages, tones are identified by their shape (contour of the fundamental frequency) and pitch range (or register) while in some African languages, tones are distinguished by their relative pitch levels (Akinlabi and Liberman, 2001), as a result tones cannot be universally applied to speech pattern classification (Chen et al., 2016). Classical ASR systems are based on context dependent tri-phone acoustic modeling and commonly use phone features, such as Mel-filtered cepstrum coefficient (MFCCs) as input features. This model and representation work well for phone recognition, but do not carry information about tone. Another challenge, is the segmentation of sentences of tonal language into words. In the SY writing and speaking system, the basic unit is syllable and not word. Consequently, the design and implementation of Multi-class Support Vector Machine in the recognition of SY context dependent tone is presented in this paper to engender and provide arguments for the use of context dependent tone segment for SY ASR. In language such as SY, tones are associated with syllable (Yang and Zhang, 2018). SY has seven possible syllable structures, these include consonant-vowel $CV$, $CVn$, digraph-vowel nasal $DVn$, digraph-vowel $DV$, vowel $V$, vowel nasal $Vn$ and syllabic nasal $n$. SY has three lexical tones: high, low and mid. In recent times, several models have been proposed for tone language ASR. These techniques can be categorized into two main classes: (i) rule-based and (ii) data-based approach. The implementation of the rule-based
scheme requires eliciting of rule-sets from knowledgeable experts. A drawback of this scheme, is the generation, organization and representation of the interdependency of the rule-set as well as unavailability of domain experts. These setbacks inspired the use of the data-driven techniques to ASR (Kumalalo et al., 2010). The most commonly used generative models for tonal language ASR are: (i) embedded (Chen et al., 2014) and (ii) explicit (Kristine, 2017; Li et al., 2016) approaches. In the embedded scheme, tone recognition is based on a multi-stream HMM decoding while in the explicit scheme, syllables within an utterance are identified first via force alignment of HMM and then tone recognition is then performed on each segmented syllable using Gaussian Mixture Model (GMM). Compared with embedded tone modeling, the explicit tone modeling approach is capable of exploiting the supra-segmental nature of the tones. There are two major approaches to explicit tone modeling: sequence based tone modeling and segment based tone modeling (Chen et al., 2014). Due the fact that articulation of human is sequential and output of pitch related feature extraction is frame based, modeling of tones using sequential model is logical. Examples of sequenced model includes the hidden Markov model (HMM) and hidden conditional random fields (HCRF), etc. A major weakness of sequenced model, is that is challenging for the sequence based models to use segment based information from contextual tones. Hence, considerable efforts are required to utilize pitch information of CD syllable. Discriminative training models such as Gaussian mixture model (GMM), support vector machines, neural network and deep network etc. are alternatives approach to sequence based model. Lately, MSVM have successfully been applied to many different speech recognition application, such as speaker verification, emotion and text classifications (N. Yang et al., 2017). Aida-zade et al., (2016) implemented a speech recognition system using SVM. In the work, SVM was used to make decisions at frame-level, and a Token Passing algorithm to obtain the chain of recognized words. TombaloGlü and Erdem (2017) developed SVM based recognizer, MFCC features of Turkish speech were extracted and SVM based classifier alongside a new text comparison algorithm was explored. The text comparison algorithm uses phoneme sequence to measure words similarity. Frihia and Bahi (2017), comparison algorithm uses phoneme sequence to text comparison algorithm was explored. The text extracted and SVM based classifier alongside a new recognizer, MFCC features of Turkish speech were used. So the conversion of frame based pitch-related observation sequence into a fixed dimensional vector. Second, the number of CD tri-tone are limited thus, reducing model confusability when compared to CD tri-phone which requires a lot of hours of segmented and labelled speech unit. Third, the availability of free software and tools for modeling and implementing MSVM. Hence, the objective of this paper is to develop a tri-tone acoustic model and explore the use sub-segmental features for SY CD tone identification.

MATERIALS AND METHOD

The Standard Yoruba context dependent tone identification problem is composed mainly of 2 steps: (1) MSVM model formulation (2) Implementation (training and testing).

Multiclass SVM Model Formulation: Given training data $T(x_1,y_1),\ldots,(x_l,y_l)$, where $x_i \in R^n$, $i = 1, \ldots, l$ and $y_i \in \{1, \ldots, k\}$ is the class of $x_i$. Solve quadratic programming problem.

$$\min_{w,\xi} \frac{1}{2} \sum_{i=1}^{n} (y_i - \langle w, x_i \rangle)^2 + C \sum_{i=1}^{k} \xi_i$$

Subject to

(C01) $y_i (\langle w, x_i \rangle + b) \geq 1 - \xi_i \quad \forall y_i \neq i$
(C02) $y_i (\langle w, x_i \rangle + b) \leq 1 + \xi_i \quad \forall y_i \neq i$
(C03) $\xi_i \geq 0 \quad \forall j = 1, \ldots, l$
(C04) $i, j, k, l \geq 0 \quad \forall i, j, k, l$

The bi-objective formulation is presented in Eqn. 1. Inequalities in Eqn.2, 3, and 5 are constraints. In the model, $w^i$ represents the hyperplane, $C$ is penalty parameter for error on the training sample, $\xi_i$ is the slack variable, $i, j, k, l$ are decision variables, $y_i$ is the class label, $b^i$ is un-regularized bias term and $w^i x$ is the decision function. Linear parameterization of

Eqns. 1 – 5 can thus, be expressed as

$$\min_{w_1,\ldots, w_k, \xi} \frac{1}{2} \sum_{i=1}^{n} (y_i - \langle w_i, x_i \rangle)^2 + C \sum_{i=1}^{k} \xi_i$$

Subject to:

$$(w_i x_i - y_i)^2 x_i \geq e_i - \xi_i \quad \text{if} \quad e_i = 1 - \delta_i; \quad i = 1, \ldots, n,$$
$$l = 1, \ldots, L$$
Based on Lagrange multipliers, Eqns. 6 and 7 are represented in dual forms as Eqns. 8 and 9.

\[
\max W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j
\]

Subject to:

\[
\alpha_i \geq 0, \sum_{i=1}^{n} \alpha_i y_j = 0
\]

(8) (9)

\( \alpha_i \) are the non-zero support vectors that determines the hyper-plane. To optimize Eqn. 6, requires the selection of suitable decision kernel. In this paper, radial basis function (RBF) kernel is used. The tunable parameters such as \( C \) and kernel parameters were selected through a combination of resampling techniques and a separate validation set.

Implementation: The Rule Based Corpus Optimization Model (RBCOM) SY database in (Sosimi et al., 2015) is used to evaluate MSVM performance utilizing the explicit tone feature. RBCOM is a corpus of SY read speech. It contains recordings from 100 native SY speakers comprising of 50 males and 50 females each producing approximately 200 utterances. The default encoding was set at 16 kHz Microsoft wave files 16bits mono. The resulting usable corpus amounted to 14,482 utterances from 98 speakers, each having between 115 to 140 single-phrase utterances. The database was divided into two partitions: 85% of data set constitute training set while the remaining 15% represents the test set.

Training and Testing: To train the MSVM, grapheme to phoneme conversion was done based on a description of the SY orthography followed by the implementation syllabification algorithm and automatic phonemic alignment of the audio using speech phonetization alignment and syllabification (SPPAS), treating each speaker’s utterances independently. The tone introduces the requirements of generation of tone annotations from SY syllable tier. The algorithm for the production of equivalent tone transcriptions is described in this section. The algorithm was implemented based on a description of the SY orthography with each vowel bearing any of three (3) distinct tones: high (H), low (L) or mid (M) and eighteen consonants (0). The pseudo code of the algorithm is presented below. MATLAB R2013a software was used for computer implementation of the algorithm. This is followed by extracting and selecting the best feature subset for predicting the class label.

\[T(i) = \text{H} \quad \text{End} = |H|\]
\[\text{For} \ m = 1 \text{ to } |H| \]
\[\text{Control} = H(m)\]
\[\text{Next} m\]
\[\text{Elseif} \ i = 2 \]
\[T(i) = \text{L} \quad \text{End} = |L|\]
\[\text{For} \ m = 1 \text{ to } |L| \]
\[\text{Control} = L(m)\]
\[\text{Next} m\]
\[\text{Elseif} \ i = 3 \]
\[T(i) = \text{M} \quad \text{End} = |M|\]
\[\text{For} \ m = 1 \text{ to } |M| \]
\[\text{Control} = M(m)\]
\[\text{Next} m\]
\[\text{Else} \quad T(i) = \text{0} \quad \text{End} = |0|\]
\[\text{For} \ m = 1 \text{ to } |0| \]
\[\text{Control} = 0(m)\]
\[\text{Next} m\]

Document_Flag = 1, j = 0;
/* Document_Flag is a binary value (0 or 1) to show when a document is closed or opened.*/
Do while Document_flag = 1
j = j + 1, k = 0 /* line number in the document*/
Do until k > Line_End (j)
/* Line_End (j) depicts the total number of characters*/
K = k + 1 /* Character number in the document*/
Read Char (j,k)
For n = 1 to End_CA
if Char (j,k) = Control (n)
Char (j,k) = T(i)
Else
End
if
Next n
End
Do
if j = End_of_Line; Document_Flag = 0
End
Do
Next i

Pitch \( f_0 \) information associated with the context dependent syllable were extracted using the short time Amplitude Magnified Difference Function (AMDF). The pitch extracted is affected by interaction between voiced and unvoiced phones, the \( f_0 \) sequence is refined using least square regression (LSR), cubic spline and

\( L_0 \)
dynamic programming (DP) (Chen and Jang, 2008) to derive the Least Square, Cubic Spline and Dynamic Programming Embedded tone features respectively. POLYFIT MATLAB function was used to obtain the slope and intercept of \( f_0 \) over each segment.

For a wider diversity CD tri-tone and contextual influence of tones, the tri-syllable unit is shifted by a syllable and the refined features are normalized using the following normalization scheme:

\[
f_{ok}(t) = \frac{f_{ok}(t) - \text{min}_{f_{ok}}}{\text{max}_{f_{ok}} - \text{min}_{f_{ok}}}
\]

Using logarithm of \( f_0 \) value and normalizing this logarithmic value of \( f_{ok} \) by min and max of each cluster \( \text{NORM}_\text{LOG}_\text{f}_0 \).

\[
f_{ok}^2(t) = \frac{\log f_{ok}(t) - \text{min}_{\log f_{ok}}}{\text{max}_{\log f_{ok}} - \text{min}_{\log f_{ok}}}
\]

\[
f_{ok}^3(t) = f_{ok}(t)/\bar{f}_{ok}
\]

In this study, a hybrid normalization scheme \( f_{tk} \) as presented in Eqn.13 is also explored (where \( k \) represents slope or intercept vector). Resulting to four dimensional feature vectors namely absolute slope, absolute intercept, normalized slope and normalized intercept, where \( w_l \) is the weight representing the contribution of each feature.

\[
f_{tk} = w_1f_{ok} + w_2f_{ok}^2 + w_3f_{ok}^3
\]

Subject to

\[
\sum_{i=1}^{5} w_i = 1
\]

The space SY CD tri-tone is determined using the expression below.

Dummy start \( + \eta \) Dummy end

Where \( \eta \) is the number of tones which in the case of SY is 3, while \( r \) is number of items to be chosen which for tri-tones is 3. For SY language has 27 distinct CD tri-tone clusters plus two (2) dummy cluster resulting in a total of 29. Having created the feature vectors, each Tri-tone context is clustered based on respective signature as presented in Figure 2.

![Fig. 1. SY Context Dependent tri-tone Tree](image)

For an unbiased estimation of training algorithm, a combination of the hold-out sampling and whole data set is explored. Repeated \( k \)-fold cross validation is performed on \( S \). The \( k \) is varied over \( n \)-number of runs (i.e. \( k = 2, 4, 6, 8, 10, 15, 20, 50 \) and \( n = 10, 20, 50, 100, 150, 200, 500, 1000 \)) in order to determine the optimal values of \( \mu_k, a_2, \gamma \) (where i.e. \( \mu_k \) is the mean value taken over all possible \( k \)-fold cross validations over \( S \), \( a_2 \) is the conditional prediction error and \( \gamma \) the mean accuracy of \( L(S') \) on \( P \), taken over all training set \( S' \) of size \( (k-1)/k|S| \) which generates the best accuracy. Having learnt the optimal model vis-a-vis \( k, n, \mu_{opt}, a_{opt}, \gamma_{opt} \) that best describes \( S \), where \( \mu_{opt}, a_{opt} \) and \( \gamma_{opt} \) are the mean taken over all possible \( k \), condition prediction error and mean accuracy segment respectively. \( E \) is used to estimate accuracy of the model using the mathematical formula presented in Eqn. 15.

\[
\% \text{Acc} = \frac{\text{Correctly Classified} E}{\text{Total of population of } E} \times 100
\]

Computer implementation of the Multiclass SVM Learning Algorithm was done using MATLAB.

**RESULTS AND DISCUSSION**

In order to capture information between the voiced and unvoiced speech segment, pitch contour refinement

*SOSIMI, AA.; ADEGBOLA, T; FAKINLEDE, OA.*
schemes are implemented, a sample of results of refining the broken pitch segment is illustrated in Figure 4.

**Baseline Performance:** The HTK baseline system was run without adaptations during the training and decoding to evaluate the proposed model. The results of the baseline MFCC and pitch system are illustrated in Figure 4.

From Figure 4, the baseline system via the DP refinement has the highest accuracy of 37.32%. On the other hand, the Least Square Embedded Tone LSET baseline system recorded an accuracy of 34.62% and the CS resulted in an accuracy of 31.52%. The highest accuracy, over all the schemes was obtained at Insertion Log Probability (ILP) between -18 and -15. The percentage of correctly recognized word increases as ILP increases, while the CD tri-phone model recorded 44.57% best accuracy. For the Multiclass SVM DP baseline, best classification accuracy was recorded at experiment setup of 200 runs and a 10 folds cross-validation as shown in Table 1. At this experimental setup, 87.9252% accuracy was recorded as illustrated in Figures 4 and 5.

Having learnt the optimal model parameters (i.e., $k, n, \mu_{opt}, \alpha_{opt} & \gamma_{opt}$) that best describes the training set, the model was evaluated with the test data. At optimal parameter settings a 97.5678% tri-tone accuracy was obtained, the confusion matrix and classification spectrum are presented in Figure 6. From the results, utilization of tone acoustic model and pitch features have shown to be effective in tone classification.

**Table 1: Summary of Multi-Class SVM CD tone Training**

| Number of Folds | Accuracy (%) |
|-----------------|--------------|
| Runs 10         | 10           | 20 | 50 | 100 | 150 | 200 | 500 | 1000 |
| 2               | 59.3145      | 59.8214 | 60.421 | 62.7105 | 63.0011 | 63.8182 | 64.2118 | 65.7102 |
| 3               | 62.5065      | 63.0113 | 63.8112 | 64.0001 | 64.2173 | 64.5276 | 65.55    | 65.819 |
| 4               | 63.9979      | 64.9901 | 65.2017 | 65.4171 | 65.4192 | 65.7125 | 65.819   |
| 5               | 65.7782      | 66.0241 | 66.3215 | 67.2022 | 68.7101 | 65.4265 | 71.6591  | 72.2125 |
| 6               | 65.5678      | 66.5155 | 66.8509 | 67.651  | 69.0011 | 85.4981 | 74.2119  | 75.5067 |
| 7               | 66.248       | 67.1105 | 68.11   | 68.2016 | 69.8273 | 85.4989 | 75.3223  | 75.89   |
| 8               | 66.235       | 67.2105 | 68.0512 | 68.3011 | 72.4501 | 80.6777 | 81.4105  | 83.3433 |
| 10              | 66.9545      | 67.3178 | 70.0031 | 86.2769 | 75.2203 | 87.9252 | 84.5117  | 86.6071 |
| 15              | 67.5432      | 72.2155 | 73.3417 | 85.4001 | 83.9143 | 87.8215 | 86.2245  | 87.3888 |
| 20              | 67.6317      | 72.5275 | 75.8341 | 85.9759 | 87.2057 | 87.8598 | 86.5152  | 86.4819 |
| 50              | 67.5432      | 72.5275 | 75.9982 | 85.1983 | 86.3671 | 86.6301 | 86.5152  | 86.4821 |

SOSIMI, AA.; ADEGBOLA, T; FAKINLEDE, OA.
On comparing the performance of MSVM and HMM for CD tone classification, the MSVM yielded the best classification accuracies.

**Conclusions:** A Standard Yorùbá Context Dependent Tone identification using Multi-class Support Vector Machine (M-SVM) have been presented in this paper. The results led to three major conclusions: SY CD tone recognition problem can be implemented with MSVM and HMM. The accuracy rates achieved using the MSVM was found to be higher than that of the HMM on the validation data sets. However, the performance of MSVM in modelling time sequential nature of continuous utterance have not been reported. In addition, its ability to handle dialectic variations which are essential characteristics of SY language is yet to be determined.

**REFERENCES**

Aida-zade, K; Xocayev, A; Rustamov, S (2016). Speech recognition using Support Vector Machines. In *Application of Information and Communication Technologies (AICT), 2016 IEEE 10th International Conference on* (pp. 1-4). IEEE.

Akinlabi, A., & Liberman, M. (2001). Tonal complexes and tonal alignment. In *PROCEEDINGS-NELS* (Vol. 31, No. 1, pp. 1-20).

Chen, JC; Jang, SR (2008). Trues: Tone recognition using extended segments. *ACM Transactions on Asian Language Information Processing (TALIP)*, 7(3), 10.

Chen, M; Yang, Z; Liu, W (2014). Deep neural networks for Mandarin tone recognition. In *Neural Networks (IJCNN), 2014 International Joint Conference on* (pp. 1154-1158). IEEE.

Chen, R. Bunescu, L. Xu, C. Liu (2016). Tone Classification in Mandarin Chinese Using Convolutional Neural Networks. In *Proc. INTERSPEECH 2016*, 2150-2154.

Eme, C. A., & Uba, E. D. (2016). A contrastive study of the phonology of Igbo and Yorùbá. *UJAH: Unizik Journal of Arts and Humanities*, 17(1), 65-84.

Frihia, H; Bahi, H (2017). HMM/SVM segmentation and labelling of Arabic speech for speech recognition applications. *International Journal of Speech Technology*, 20(3), 563-573.

Kristine, MY (2017). The role of time in phonetic spaces: Temporal resolution in Cantonese tone perception. *Journal of Phonetics*, 65, 126-144.

Kumolalo, FO; Adagunodo, ER; Odejobi, OA (2010). Development of a Syllabicator for Yorùbá language. *Proceedings of Obafemi Awolowo University TekConf*, 47-51.

Li, WS; Iniscalchi, SM; Chen, NF; Lee, CH (2016, December). Using tone-based extended recognition network to detect non-native Mandarin tone mispronunciations. In *Signal and Information Processing Association Annual
Sosimi, A; Adegbola, T; Fakinlede, O (2015). A Supervised Phrase Selection Strategy for Phonetically Balanced Standard Yorùbá Corpus. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 565-582). Springer, Cham.

Tombalolu, B; Erdem, H (2017). A SVM based speech to text converter for Turkish language. In *Signal Processing and Communications Applications Conference (SIU)*, 2017 (pp. 1-4). IEEE.

Yang, L.; Xie, Y.; Zhang, J. (2018) Improving Mandarin Tone Recognition Using Convolutional Bidirectional Long Short-Term Memory with Attention. Proc. *INTERSPEECH* 2018, 352-356.

Yang, N.; Yuan, J.; Zhou, Y.; Demirkol, I.; Duan, Z.; Heinzelman, W.; Sturges-Apple, M. (2017). Enhanced multiclass SVM with thresholding fusion for speech-based emotion classification. *International Journal of Speech Technology*, 20(1), 27-41.