PRINT ISSN 1119-8362
Electronic ISSN 1119-8362

Full-text Available Online at
https://www.ajol.info/index.php/jasem
http://ww.bioline.org.br/ja

J. Appl. Sci. Environ. Manage.
Vol. 23 (5) 895-901 May 2019

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, Àkọ̀kà, 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

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

**Corresponding Author Email: asosimi@unilag.edu.ng*