# Computational Study of Fundamental Frequency of Standard Yorùbá Bi-Syllabic Vocalizations

**Michael Adébísí Fáyẹ̀míwò and Ọdẹ́túnjí Àjàdí Ọdẹ́jọ́bí**

Department of Computer Science and Engineering,
Faculty of Technology, Obafemi Awolowo University,
Ile-Ife, Osun State, Nigeria;
*mfayemiwo@gmail.com*
*oodejobi@oauife.edu.ng*

**Abstract** This paper presents the development and evaluation of a computational model for the fundamental frequency ($F_0$) of Standard *Yorùbá* (SY) bi-syllabic vocalizations. This was done with a view to approximating the $F_0$ curves on SY syllables in the context of speech applications and other speech technologies.

A list of 117 SY syllables which are the most frequently occurring syllables in *Yorùbá* newspapers and textbooks was compiled and the speech sound corresponding to the selected SY syllables were recorded for 5 adult native male speakers of SY. The $F_0$ of the speech data were extracted using Praat speech processing package. Thereafter, Least Square Method (using polynomial degree of 1 to 7) was used to design the computational model for the $F_0$ patterns extracted. The computational model was evaluated using the qualitative technique.

The results of the model showed that polynomials starting with degree 4 gave a good approximation for bi-syllabic vocalizations. The study thereafter established the computational and perceptual correlations between the $F_0$ curves and the three *Yorùbá* tones (High, Mid and Low) for SY bi-syllabic vocalizations. The modeling of $F_0$ contour for *Yorùbá* tones for words and continuous speech are the areas of further research works, in which the principle of this work could be extended.

***Keywords:*** Fundamental Frequency; Standard *Yorùbá*; *Yorùbá* tones; $F_0$ curves; Synthesizing $F_0$ pattern.

## 1. INTRODUCTION

The *Yorùbá* alphabet consists of 25 letters and uses the familiar Latin characters. The letters are made up of 18 consonants (b, d, f, g, gb, h, j, k, l, m, n, p, r, s, ṣ, t, w, y) and seven oral vowels (a, e, ẹ, i, o, ọ, u). The consonant 'gb' is a digraph, i.e. a consonant represented with two letters. There are five nasalized vowels in the language (an, en, in, ọn, un) and two pure syllabic nasals (m, n). SY has three phonologically contrastive tones; a high tone represented with an acute accent mark ( ´ ), a mid-tone that is usually left unmarked but in certain circumstances marked with macron (-) and a low tone represented with a grave accent mark ( ` ).

A *Yorùbá* syllable with a different tone has a different lexical meaning. Thus to completely recognize a spoken *Yorùbá* syllable, a speech recognition system needs to recognize a base syllable but also correctly identify the tone. Hence, tone classification of *Yorùbá* speech is an essential part of a *Yorùbá* speech recognition system.

Received May, 24, Reviewed May 30, 2014

Speech has been the principal form of human communication since it began to evolve (Dikshit, 2004). The rate of vibration of the cords is called fundamental frequency ($F_0$). In human speech production, the vocal chords vibrate at a temporal frequency to produce a semi-periodic air flow through the vocal tract and this frequency is the $F_0$ of the output speech signal (Chomphan, 2012b). It is an essential feature among other speech features, which carry prosodic information of the natural speech. Therefore, in the modern speech technology, e.g., speech recognition, speech analysis and synthesis, it is necessary to model the $F_0$ with high accuracy.

The main perceptual attribute of $F_0$ is the pitch, which actually is an overall perceived spectral quality. Studies in the literature on tone languages e.g. Thai (Chomphan, 2011) and (Chomphan, 2012a), Mandarin (Liu et al., 1998) and (Keatinga and Kuo, 2012), and Cantonese (Gu et al., 2011) have shown that tone is an essential feature for a speech unit of syllable. For a tonal language like *Yorùbá*, tone is an important component of sound because words with the same phoneme sequences may have different meanings if they have

different tones, for example, owó (Money) and òwò (Trade). Therefore, tone is one of the most important factors in the SY speech research field in order to make a system that has high intelligibility and naturalness. The tone is indicated by contrasting variations in contour of $F_0$ at the syllabic level. *Yorùbá* has three lexical tones named: mid (M), low (L), and high (H) and this tone is correlated to $F_0$ pattern on syllables.

This study therefore designed a computational model that best approximate the acoustic correlates of the pitch contour $F_0$ for Standard *Yorùbá* speech signal. The selected Standard *Yorùbá* bi-syllabic vocalizations were collected, recorded and pre-processed and the $F_0$ was extracted using the Praat speech processing software and the $F_0$ was imported into Matlab for further analysis. The design and implementation of computational model to approximate the $F_0$ extracted from the speech data collected were done using appropriate numerical computation technique (Least Square Method) in Matlab environment. The computational model was evaluated using the Qualitative (i.e., Mean Opinion Score) technique. The speech data that was considered in this research work was Standard *Yorùbá* of bi-syllabic vocalizations and only male voice was used for data in the extraction of the $F_0$.

## 2. LITERATURE REVIEW

The need for flexible speech modification methods is increasing in both commercial and scientific fields (Kawahara et al., 1999) and the method consists of a fundamental frequency ($F_0$) extraction. Xu and Wang (2001) proposed a preliminary framework for accounting for certain surface $F_0$ variations in speech and the framework consists of definitions for pitch targets and rules of their implementation. The $F_0$ of voice speech is the most important feature among all of the features known to carry prosodic information. Therefore, $F_0$ is an inherently supra-segmental feature of human speech. The $F_0$ contours of an utterance convey the stress, intonation and rhythmic structures, which determine the naturalness and intelligibility of synthetic speech (Chomphan, 2011). As a result, the appropriate modeling of $F_0$ contour plays a significant role in the Thai speech synthesis, the statistical modeling of $F_0$ contour has been conducted by Chomphan (2011) in the implementation of both speaker-dependent and speaker independent systems. Lately, the Fujisaki's model has been applied within a speaker-independent system as extended modules.

Fáyẹ̀míwò and Ọdẹ́túnjí (2014) presented a study on the development and evaluation of a computational model for the fundamental frequency ($F_0$) of Standard *Yorùbá* (SY) monosyllabic vocalizations. The study was done with a view to approximating the $F_0$ curves on SY syllables in the context of speech applications and other speech technologies, in which a list of 39 SY syllables was compiled and the speech sound corresponding to the selected SY syllables were recorded for 5 adult native male speakers of SY. The $F_0$ of the speech data were extracted using Praat speech processing package. Thereafter, Least Square Method (using polynomial degree of 1 to 7) was used to design the computational model for the $F_0$ patterns extracted. The model was later evaluated using the quantitative and the qualitative techniques. The results showed that the $7^{th}$ degree polynomial had the lowest RMSE value for monosyllabic speech data. The results of the model showed that polynomials with degree 3 gave a good approximation for monosyllabic. The study established the computational and perceptual correlations between the $F_0$ curves and the three *Yorùbá* tones (High, Mid and Low) for SY monosyllabic vocalizations.

In Chomphan (2012b), an approach of structural modeling from Mandarin Chinese was adapted to model Thai tones. The $F_0$ contour was modeled by using a structural control which consists of locating a number of normalized $F_0$ targets along time axis in logarithmic scale. The $F_0$ targets or pitch targets were specified by amplitudes and transition time. The RMS was used to evaluate the accuracy of the synthesized $F_0$ contour.

Moreover, another study has been conducted by using a structural model which is based on the assumption that the behavioural characteristics of vocal-fold elongation in vibration could be approximated by those of a simple forced vibrating system (Chomphan, 2012b). The RMS error was calculated for the evaluation of the model performance for both mentioned speech models and also for all speech styles including angry style, sad style, enjoyable style and reading style.

Daniel et al., (2014) investigated tone realisation in continuous vocalizations in *Yorùbá,* in which features influencing syllable pitch targets in continuous vocalizations in *Yorùbá* were investigated in a small speech corpus of 4 speakers. It was found that the previous syllable pitch level is strongly correlated with pitch changes between syllables and a number of approaches and features were evaluated in this context. The resulting models was used to predict utterance pitch targets for speech synthesisers.

## 3. DESCRIPTION OF DATA

The speech sounds of 5 male adult who are SY native speakers were recorded and a list of 117 SY syllables was compiled. The list comprised of 13 phonetic bi-syllabic words each with 9 possible combinations of tones, resulting in 13 × 9 = 117 bi-syllabic words. The 117 words are the most

frequently occurring words in selected texts from *Yorùbá* newspapers and textbooks.

One of the recorded voice samples was used for the actual experiment while the other four samples were used for control and confirmation. The selected syllables were recorded in quiet environments with sufficiently low background noise, thus guaranteeing a reproducible setting and the required acoustic conditions. Praat software was used to record the speech samples.

Data of the format *V-CV, V-CVn, N-CVn, CV-CV, N-CV, CVnCVn* and *CV-CVn* were collected. **Table 1** shows the syllable types and examples of collected data that were used in this study.

## 4. MODEL DEVELOPMENT

The computational versatility and simplicity of the least square method motivated us to apply it in approximating the $F_0$ data. The program for the least square approximation was written as an M-File in Matlab environment. By using this model, we approximated the $F_0$ extracted for the speech samples using polynomial of degree 1 to 7.

Table 1: Types and Sound Samples Collected

| Type | Sample | Total Collected |
|---|---|---|
| V-CV | *ade, ọgba* | 18 |
| V-CVn | *adun, awọn* | 18 |
| N-CVn | *Nkan* | 9 |
| CV-CV | *pade, giga* | 18 |
| N-CV | *nlọ, nbẹ* | 18 |
| CVn-CVn | *kankan, dundun* | 18 |
| CV-CVn | *gbadun, dundun* | 18 |
| Total | | 117 |

## 5. SYSTEM EVALUATION

The original values of $F_0$ was replaced by the approximated values of $F_0$ in the speech data. The original speech data was initially manipulated and we extracted the pitch tier from it.

In the qualitative evaluation, listeners were asked to rate the quality of speech sound after the original $F_0$ curve had been replaced by the approximated $F_0$. There are two aspects to this qualitative evaluation:

I. The first test is the intelligibility test, where each syllable with modified $F_0$ curves were played and the listeners were asked to write down what they heard.
II. The second test is the naturalness test; the listeners were presented with the speech samples of the original and modified $F_0$ curve at random. They were asked to rate the quality of the sample in terms of how close they were when compared to those produced by humans.

The evaluation was performed on stylization functions corresponding to the 1st, 2nd, 3rd, 4th, 5th, 6th and 7th degree polynomials. Nine adult native speakers of SY in Obafemi Awolowo University Ile-Ife (students with age ranging between 19 and 35 years) were invited. To ascertain their understanding of SY, some recorded natural speech sounds were played to them and they were asked to write down what they heard. Those who were unable to produce 100% accuracy in this test were excluded from the evaluation process. In the end, five of the speakers participated in the qualitative evaluation. For the intelligibility test, 35 bi-syllabic words with the approximated $F_0$ curves were played to the listeners. The syllables were selected to reflect phonetic balance in the tests data and the volunteers had no prior knowledge of the syllables that were played. After a speech sound was played, the listeners were asked to write down what he or she heard. The results for 1st, 2nd, 3rd, 4th, 5th, 6th and 7th degree polynomials are as recorded in **Table 2**.

To conduct the naturalness test, the bi-syllabic words with the original $F_0$ curve and the corresponding ones with modified $F_0$ curve were played at random. In this test, the listeners were asked to rank the quality of what they heard on a scale of 1 (very poor) to 5 (very good). The scale is as shown in **Table 3**.

## 6. RESULTS

The result from the naturalness test is shown in **Table 4**. In the evaluation result as presented in **Table 4**, degree 1 and 2 have mean of 24% and 28% respectively and they are both ranked as very poor in the ranking scale. Degree 3 had mean of 48% which is also ranked as poor. This shows that degree 1, 2 and 3 gave poor result for approximation of the extracted $F_0$ values. Degree 4 had mean 80% and this is ranked as good in the ranking scale. However, degree 5, 6 and 7 had 96%, 100% and 100% respectively and they are all ranked as very good. This result shows that a good approximation of the $F_0$ of bi-syllabic words is only achievable with polynomial degree 4 and above.

## 7. CONCLUSION

This paper has presented the development and evaluation of a computational model for the fundamental frequency ($F_0$) of Standard *Yorùbá* (SY) bi-syllabic vocalizations. The results have shown that a 4th degree is adequate for modeling the $F_0$ curves on *Yorùbá* bi-syllabic speech data. We are therefore able to establish the computational and perceptual correlations between the $F_0$ curves and the three *Yorùbá* tones for SY bi-syllabic vocalizations. The result of this research will provide a better understanding of the characteristic of $F_0$ of SY vocalizations and the model developed can serve as a resource for *Yorùbá* speech recognition systems, text-to-speech systems and other speech related applications. However, modeling of $F_0$ contour for *Yorùbá* tones for words and continuous speech are the area of further research work, in which the principle of this work could be extended.

Table 2: Mean % Intelligibility of SY Syllables with Stylized $F_0$ Curve

| | degree1 | Degree2 | Degree3 | Degree4 | Degree5 | Degree6 | Degree7 |
|---|---|---|---|---|---|---|---|
| Bi-syllabic | 24 | 40 | 76 | 100 | 100 | 100 | 100 |

Table 3: Rating Scale for MOS

| Rating | Score |
|---|---|
| Very poor | 1 |
| Poor | 2 |
| Average | 3 |
| Good | 4 |
| Very Good | 5 |

Table 4: Result of naturalness test for '*ìgbà*'

| Degree | Listener | | | | | Mean % | Rating |
|---|---|---|---|---|---|---|---|
| | L1 | L2 | L3 | L4 | L5 | | |
| 1 | 1 | 1 | 2 | 1 | 1 | 24 | 1 |
| 2 | 1 | 2 | 2 | 1 | 1 | 28 | 1 |
| 3 | 2 | 2 | 3 | 3 | 2 | 48 | 2 |
| 4 | 3 | 4 | 5 | 4 | 4 | 80 | 4 |
| 5 | 4 | 5 | 5 | 5 | 5 | 96 | 5 |
| 6 | 5 | 5 | 5 | 5 | 5 | 100 | 5 |
| 7 | 5 | 5 | 5 | 5 | 5 | 100 | 5 |


## ACKNOWLEDGEMENTS

We would like to thank the people of Computing and Intelligent Systems Research Group (CISRG), in the Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria, for the invaluable support towards the successful completion of this research.


## REFERENCES


[1] Chomphan, S., "Modeling of Fundamental Frequency Contour of Thai Expressive Speech using Fujisaki's Model and Structural Model. Journal of Computer Science," 7 (8):1310–1317, 2011.

[2] Chomphan, S A., "Control of Fundamental Frequency Contour for Hidden Markov Model-Based Thai Speech Synthesis. American Journal of Applied Sciences," 9 (2):259–264, 2012.

[3] Chomphan, S., "Structural Modeling of Fundamental Frequency contour for Thai Tones," American Journal of Applied Sciences, 9 (10):1736–1741, 2012.

[4] Daniel R., Niekerka, V., Barnard, E., "Predicting utterance pitch targets in Yorùbá for tone realisation in speech synthesis," Speech Communication (Elsevier). Volume 56, Pages 229–242, 2014.

[5] Fáyẹ̀míwò M.A., Ọdẹ́jọbí O.A., "Computational study of fundamental frequency of standard Yorùbá monosyllabic vocalizations," International Journal of Innovative Research in Science, Engineering and

Technology. ISSN: 2319-8753. Vol 3 (4). pp 11621-11629, 2014.

[6] Gu, W., Fujisaki, H., Hirose, K., "Analysis of Fundamental Frequency Contours of Cantonese Based on a Command-Response Model. SSW5," The University of Tokyo, Shanghai Jiaotong University, pp. 1069–1072, 2011.

[7] Hombert, J.M., "Consonant Types, Vowel Height and Tone in Yoruba," Studies in African Linguistics, 8(2):173–190, 1977.

[8] Kawahara, H., Katsuse, I.M., Cheveign, A., "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency based F0 extraction: Possible role of a repetitive structure in sounds," Speech Communication (Elsevier). 27, pp 187-207, 1999.

[9] Keating, P., Kuo, G., "Comparison of Speaking Fundamental Frequency in English and Mandarin," The Journal of the Acoustical Society of America, 132 (2), 1050, 2012.

[10] Liu, S., Doyle, S., Morris, A., Ehsani, F., "The Effect of Fundamental Frequency on Mandarin Speech Recognition," In Proc. of ICSLP, volume 6, pages 2647–2650, Sydney, Australia, 1998.

[11] Xu, Y., Wang, Q.E., "Pitch targets and their realization: Evidence from Mandarin Chinese," Speech Communication (Elsevier). 33 (4), pp. 319-337, 2001.

# Biography

Dr. Odejobi's research is in the area of Computing and Intelligent Systems Engineering with focus on speech and language engineering. Dr. Odejobi was a Visiting Scholar to the School of Information Technology and Electrical Engineering at the University of Queensland, Brisbane, Australia as a fellow of the United Nations University International Institute of Software Technology (UNU-IIST). He was a Commonwealth Scholar (British) at the University of Aston in Birmingham, United Kingdom. Dr. Odejobi also had visiting scholar positions at the Phonetics Laboratory of the University of Oxford, Oxford, England and The Centre for Speech Technology Research (CSTR), of the University of Edinburgh, in Scotland, United Kingdom. Dr. Odejobi was a Marie Curie Research Fellow on the Constraint Reasoning Extended to Enhance Decision (CREED) Project at the Cork Constraint Computation Centre (4C), of the University College Cork (UCC), Cork, Ireland (The Republic). He was a Research Academic in the same institution between 2008 and 2009. Dr. Odejobi is a consultant to a number of National and International organisations. For example, he served as a consultant to the African Languages Technologies Initiatives (Alt-I) on the Microsoft Vista Operating System Localisation Project. He is currently a Senior Lecturer in the Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria.

Fayemiwo Michael's research area is in Computational speech recognition. He is a member of Computing and Intelligent Systems Research Group, Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife. He is a graduate of Computer Science, Federal University of Agriculture, Abeokuta (FUNAAB), Ogun State and was awarded B.Sc (Honours) Computer Science in January 2009. He had his post-graduate study at Obafemi Awolowo University, Ile-Ife, Osun State, and was awarded M.Sc. (Honours) Computer Science in May, 2013. Fayemiwo is presently a PhD student in Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Osun State, Nigeria. He is currently an Assistant Lecturer in the Department of Computer Science, Oduduwa University, Ipetumodu, Ile-Ife, Osun State, Nigeria.