Automatic detection of the most rhythmic speaking style and context to approximate stress-timed rhythm for Non-native Learners (L2) of Arabic Language

Bader Matar F Alotaibi
Department of Computer Science
University of Sheffield
Sheffield, UK
bmfalotaibi1@sheffield.ac.uk

Roger K Moore
Department of Computer Science
University of Sheffield
Sheffield, UK
r.k.moore@sheffield.ac.uk

Abstract

The role of rhythm and entrainment in learning interaction of L2 still is not well understood. Previous studies in this field have used particular contexts such as drummed speech, music and sentences from normal speech (such as read speech). In addition, previous studies used rhythm metrics based on segmental durations to measure the variation of rhythmicity of L2 instead of rhythm metrics based on envelope spectral analysis. However, consequent studies have used some spectral analysis to detect certain specific patterns of speech rhythm rather than the entirety of the rhythmicity. Finally, Prosodic features (such as rhythm and intonation) of native and non-native speakers of English have been evaluated using an automated scoring system. This study aims to find whether measured and scored rhythmicity signals of religious and poem contexts uttered in different speaking styles and speech rate are able to be used as useful speech materials to automatically train and teach rhythmic proficiency of spoken Arabic language for L2 learners and in hence speed up the acquisition of stress-timed rhythm and the extent to which these are useful in this regard as compared to methods used by previous studies. Results showed that a trained J48 classifier with a full feature set to predict a rhythm score for every L2 utterance, using ten-fold cross-validation has achieved an accuracy of 94%. Therefore, the results of this study are expected to be a starting point for improving Speech and Language Technology in Education (SLaTE) technologies. For example, acquisition skills to approximate stress-timed rhythm for L2 learners could be developed via establishing automatic religious and poem rhythm generation models for prosodic training. Furthermore, such models can be evaluated using other stress-timed languages such as English and Russian.

Keywords: stress-timed rhythm, Prosodic, SLaTE

1 Introduction

Speech and Language in Technology Education, or SLaTE, refers to the use of computer systems to assist in speech training. These are generally automated voice-interactive systems used by educators and learners to facilitate the acquisition of second languages (Hardison, 2004). In other words, SLaTE systems are computerised language teachers that are able to diagnosis and evaluate spoken inputs of users and generate feedback (Bang et al., 2013). Such systems encourage both educators and learners to create meaningful conversational interaction by using instructional materials outside the classroom.These systems use two phonological levels for assessment: segmental and suprasegmental. A number of studies have attempted to create automated training systems for speech prosody, including intonation and rhythm, to assess and train non-native speakers. For instance, Chen and Zechner (2011) proposed an automatic approach to extract both rhythmic and non-rhythmic features from speech of non-native English learners who spoke Mandarin as a native language. These features were compared to determine which one would be the most efficient to improve the prediction of a human rating model of scoring of non-native speakers proficiency. The advantage of this study that differentiates it from

previous studies is that their approach was applied to a large non-native speech corpus containing 40 hours of audio. This corpus was partitioned into three sections according to its purpose in the study: training, evaluation and testing. Human raters were asked to score prosody aspects of reading responses of the audio such as intonation. A three-point scale was used by human raters: 3 denoted a high-level, 2 a medium-level and 1 a low-level. Another automatic assessment system of prosody for second language learning was proposed by Arias et al. (2010) The prosodic features studied here were stress and intonation, the latter of which refers to the pitch variation in spoken units and which is considered as an output of the interaction between different prosodic features such as loudness, pitch-range, tone, rhythmicity and tempo. In this study, the proposed system aimed to use intonation and energy contours together to assess the stress patterns of (Spanish) L2 learners compared to a reference of native speech of the target language (English). Reference signals of native speech and learner input signals were compared using Dynamic Time Warping to measure the distance of the similarity between the speech of a native speaker and the speech of the L2 learner and then produce an objective score that showed whether learner utterances were close or deviated from native speech. This study focused attention only on primary stress as Tepperman and Narayanan (2005) noted that misplacing primary stress has an influence on the lexical meaning. Similarly,Shahin et al. (2016)proposed an automatic method using Deep Learning for classifying lexical (primary) stresses only. However, both these studies did not quantify secondary stress. In another study, however,Ferrer et al. (2014) proposed an automatic stress detection of English sentences as pronounced by Japanese students. A Hidden Markov Model (HMM) was used in this study to determine and classify both primary and secondary sentence stress level using a predefined dictionary. If Arias et al. (2010) proposed system had adopted this approach to determine levels of stress and incorporated other speech rhythm patterns such as duration, which represents timing and speaking rate, it would have improved the richness of the data. Most recently, Sztahó et al. (2018) used two parameters to assess speech rhythm of children with impaired hearing: the lengths of time interval between successive vowels and vowel duration. These two parameters do not fully describe rhythm and are missing stress. In their study, they measured the intonation quality (compared to a reference point) of the speaker according to the direction of change in pitch of speech. The authors showed that Hungarian children with hearing impairment who used a computer-based speech prosody learning system including rhythm and intonation were able to produce a good level of prosodic features in their speaking, specifically in terms of rhythm and intonation. Moreover, the authors pointed out that their system could be adapted to other languages. Zechner et al. (2011) study utilizes two prosodic features: tone and stress for training decision tree (C4.5) classifier for scoring non-native English speech. Similarly,(Jang, 2009) study utilizes eight durational features of rhythm to measure and evaluate automatically the proficiency of Korean speakers English pronunciation. They point out that machine scoring of rhythm can be more reliable and robust by discovering F0 and intensity features. In our study we employ such features in addition to other features in order to label and annotate phrase boundaries and words prominence.

2 Prosody in SLaTE

In its totality, language is identified aurally by the following three features: (1) The acoustic properties of phonemes known as segmental features; (2) high level features of morpho-syntax and lexicon; and (3) prosody, a supra-segmental features. Speech prosody can be defined as a combination of suprasegmental elements: rhythm, which includes pauses and durations; and intonation, which includes pitch and loudness. The accuracy of these features contributes to the perception of a speaker being native or not(Zechner et al., 2011). Prosody can be detected in terms of pitch but this is not the most reliable measurement (Jang, 2009). In fact, Sundström et al. (1998) measured the similarity of pitch between native and non-native speakers. Two speech signals were used in this study: the students and the teachers. Student speech signals were labelled and aligned with the correct utterance of the teacher. Two signal features were measured in the alignment process: fundamental frequency and duration. These were modified and resynthesized to conform with the correct prosody as uttered by the teacher.

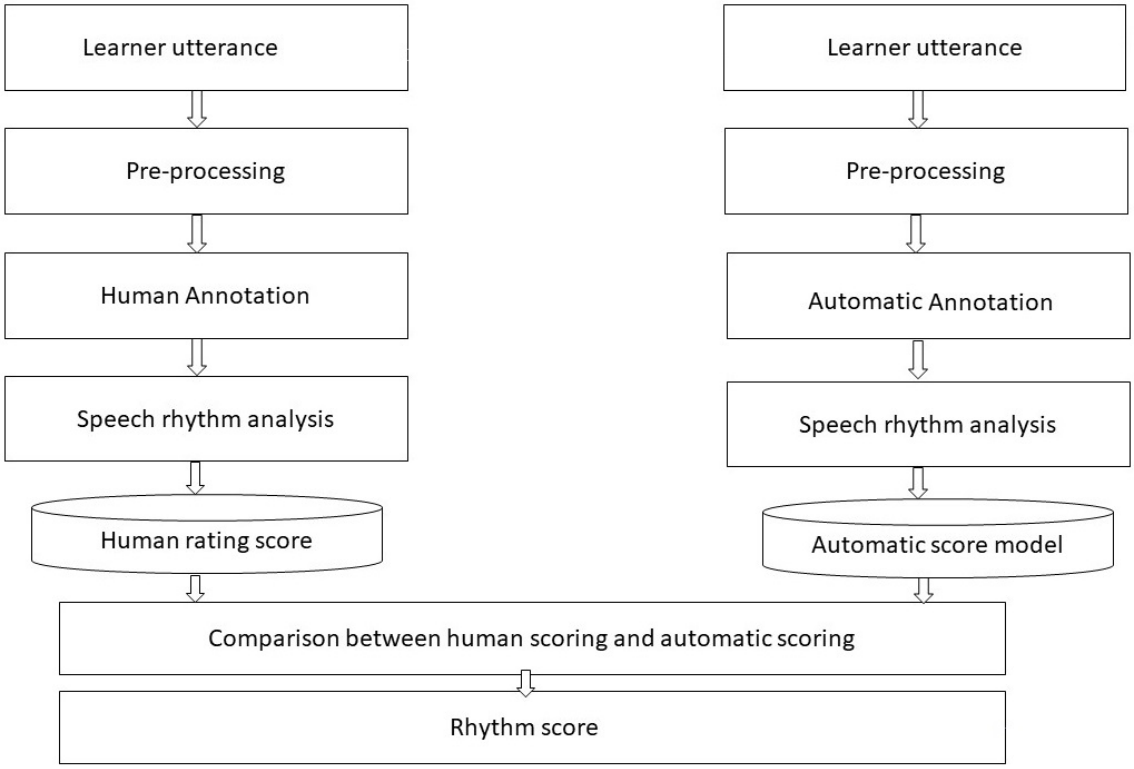

Figure 1: *Block Diagram of proposed system of speech rhythm proficiency in SLaTE for Arabic Foreign Learner (AFL)*

3 The Proposed System

The main aim of the proposed system is to determine and assess the proficiency of speech rhythm of L2 learner using particular contexts (i.e. religious texts and poems uttered in different styles and situations). The direct objective of this study is to measure and score the rhythmic proficiency of learners corresponding to each context and sub-context using the scoring system. That can be achieved by comparing two scores: human score of L2 learners utterances which serves as a gold standard and automatic score. In order to achieve this, annotated data must be collected and speech rhythm features extracted in order to train the speech rhythm classifier. The block diagram of this system is shown in figure 1. The proposed system, including speech dataset and two module: human scoring and automatic scoring of L2 speech rhythm. In the next section, the speech database and tasks of two modules are explained in detail.

4 Method

Automated labels of rhythmic patterns which were taken from rhythm perception model are used to perform a classication in Weka software Hall et al. (2009),a machine learning and data mining tool. Results of this classification were compared with the results obtained with the same classier that uses human rhythm labels of L2 utterances. In this study a parametric approach was used to describe rhythmic cues of L2 utterances. To achieve that a rhythmic tagger embedded into a Praat-based platform allows to annotate rhythmic features.

4.1 Participants

The subjects were 9 males range from between 19 and 50 years old whose native languages included those within both stress-timed and syllable-timed families. The number of participants for each language is as following: English: 3, Jamaican Patois (English creole): 1, Urdu: 1, Bangli: 1, Kurdish: 1, Farsi: 1 and Somali: 1. It should be noted that the Arabic language level of the participants was not equal, with 7 beginners and 2 intermediate participants and that none of the participants had any speaking or hearing difficulties.

4.2 Speech materials

Subjects were instructed to read aloud all the following materials:

Religious text: two Quranic verses which were transcribed as "Al-hamdu l-illa: hi rabbi l-a:a:lami:n. R-rahma:ni R-rahi:m " translated to English as "praise and thanks to Allah the Lord of mankind, jinn, and all that exists. He is the Most Gracious and the Most Merciful. In addition, one long sentence of religious text for melodic chanting was chosen for joint recitation. Three types of recitation are performed here: melodic chanting, normal chanting and joint melodic chanting.

Poem: two verses, each with a particular meter in Arabic is known as Ramal which is similar to Anapaestic in English poem. These verses were transcribed as ala a 'l-badru alayn min thaniyyti 'l-wad wajaba 'l-shukru alayn m da li-l-lhi da" they were translated to English as The full moon rose over us from the valley of Wada And it is incumbent upon us to show gratitude for as long as anyone in existence calls out to Allah. Subjects were asked to read each sentence aloud at different speech rates: slow, normal and fast. They were also asked to perform these sentences in two speaking styles; slower melodic chanting and normal recitation. Each recording was digitised at 16kHz sampling rate and stored at 16 bits per sample mono recordings in .wav format.

5 Experiment:Automatic Scoring of AFL's speech rhythm

5.1 Data

Utterances of four sentences, as in Table 1, were spoken by 15 L2 learners and two Arabic native speakers . Each sentence was uttered by each of 17 speakers including 15 non-native learners and 2 native speakers. The dataset was used to predict rhythm for any speech input and to train a speech rhythm classifier.

5.2 Pre-processing

First, the speech signal of each utterance was prepared for processing by eliminating all factors that affect the quality of the signal. Each speech signal was sampled at 16 kHz to ensure fidelity of the recording and silences at the beginning and end of the signal was removed. After this, any noise from the power supply was reduced by applying a high-pass filter.

5.2.1 Automatic Onset Detection

One of the most common methods which is used to annotate rhythm is onset-offset, i.e. marking the beginning and the ending of utterance input. Thus, modifying the rhythmic properties of any audio signal necessitates first localising the starting points of acoustics events based on transition regions and sudden burst energy. In using this process, a temporal segmentation of the input signal is yielded known as onset detection (Degara et al., 2010). Candidate onsets can be found by computing spectral difference function. Strong onset candidates can be discovered using peak picking. Consequently, speech input can be segmented into chunks between two adjacent onsets. As a result, the start of stresses in the speech input can be determined. Such onset times were used as rhythm feature to determine the onset rate. An example of detected onset times and onset strength envelope of the religious Arabic verses (native Arabic reciter of Quranic verses) are presented in figure 2.

5.3 Transcription and Data Annotation

Transcribing speech units requires an automatic speech segmentation,the process of portioning spoken utterances into chunks (Rabiner Juang 1993); (Pfeiffer 2001); (Anwar et al., 2006). A number of cues can be used by the segmentation tool to identify the boundaries of speech chunks. These include power spectral density (PSD), zero crossing rate (ZCR), and intensity and pitch (Anwar et al., 2006). For the segmentation process, two methods can be used: splitting utterances into words or splitting each utterance into beats (Lidji et al., 2011). Either method requires speech transcription. Therefore,all speech materials were transcribed and then labelled automatically with Praat (Boersma et al., 2002). The transcribed files which were thus generated of each utterance comprised seven label types: a segmentation tier of silence and utterance to show voiceless and voiced stretches of speech; a transcription tier of the whole sentence; a words tier; a syllable nucleus tier to count syllables and present their locations; a syllables tier; a syllabification tier to present manually annotated consonantal and vocalic intervals; and a beat tier to identify the beat locations. Such locations include the suprasegmental properties of speech signals (such as stress, pitch and intensity).

context main type	context sub-type	Speaking style	Sentence	Num of syllables
Religious Text	Quranic verses	Normal Individual Recitation	Al-hamdu l-illa:hi rabbi l-a:a:lami:n. R-rahma:ni R-rahi:m	49
Religious Text	Melodic Sentence	Melodic Chanting of Joint speech	Allahu Akbar Wa l-illah:hi Al-hamd	32
Oratory	Religious oratory	Normal Individual Recitation	Kaan Rasul Allah Sala Allah Alih Wa salam Kulukhu Al-quraan	45
Poem	Iambic(x /)	Melodic Chanting Recitation	Tala a 'l-badru alayn min thaniyyti 'l-wad wajaba 'l-shukru	36

Table 1: *The sentence script of speech samples for each speech context and speaking style.*

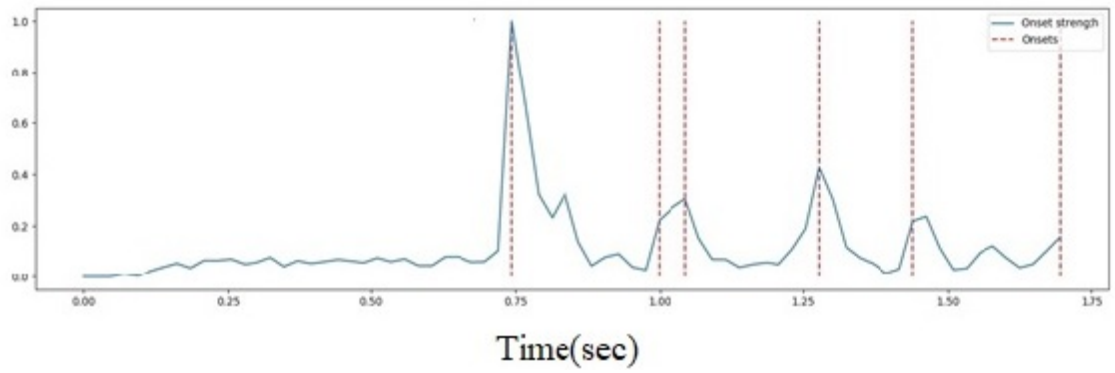

Figure 2: *onset times and onset strength envelope of the Quranic verses:"Al-hamdu l-illa:hi rabbi l-a:a:lami:n. R-rahma:ni R-rahi:m"*

5.4 Speech Rhythm Cues

The rhythm of each utterance was mainly analysed based on the detection of a number of reliable cues of speech rhythm such as syllables nucleus, phrase boundaries, intensity contour and pitch contour. Therefore, such cues have been labelled automatically in utterances segments using Pratt (De Jong and Wempe, 2009) and Essentia (Bogdanov et al., 2013). Figure 3 shows the waveform of an Arabic religious verse Al-hamdu l-illa:hi rabbi l-a:a:lami:n. R-rahma:ni R-rahi:m with the seven annotated tiers which have been mentioned previously. Similarly, figure 4 shows the waveform and annotation tiers of an Arabic poem excerpt alaa 'l-badru alayn min thaniyyti 'l-wad wajaba 'l-shukru alayn m da li-l-lhi da".

6 Human scoring

All utterances were perceptually assessed using subjective listening. Two Arabic native experts listen to each utterance and assigned a rhythm score for each one. Modelling automatic scoring system and evaluate its performance necessitate manual scoring to compare with and calibrate. Each utterance is scored in the range 3 (poor), 4 (semi-native), 5 (like native).

7 Features Extraction

Our automatic system extracted 15 features based on duration, syllable nucleus, pauses, speaking rate, average of syllable duration, phrase boundaries, prominence mean F0, standard deviation of F0, mean and standard deviation of intensity, onset times, onset rate, beat and tempo (beat per second). All extracted rhythm features were employed as metrics and then they integrated into one overall score of rhythmicity. Figures 5 shows intensity and pitch contours and relevant features sets. In the next sections, the acoustic cues of two of the most important features of speech rhythm: Phrase boundaries and prominence will be presented.

7.1 Phrase boundaries

Phrase boundaries (PPh) in the speech signal can be determined by one or combination of the acoustics cues: final rising of pitch contour, sharp fall in intensity and lengthening of the last word (Dominguez et al., 2016).

7.2 Prominence

Each PPh contains at least one word that can be labelled as prominent (Dominguez et al., 2016). Prominent words are cued by one or more of acoustics characteristics such as longer duration, high intensity and F0 Peak (Xu, 2013).

7.3 Automatic Beat Detection

Vocal rhythm is defined as groups of beats which include the suprasegmental properties of speech

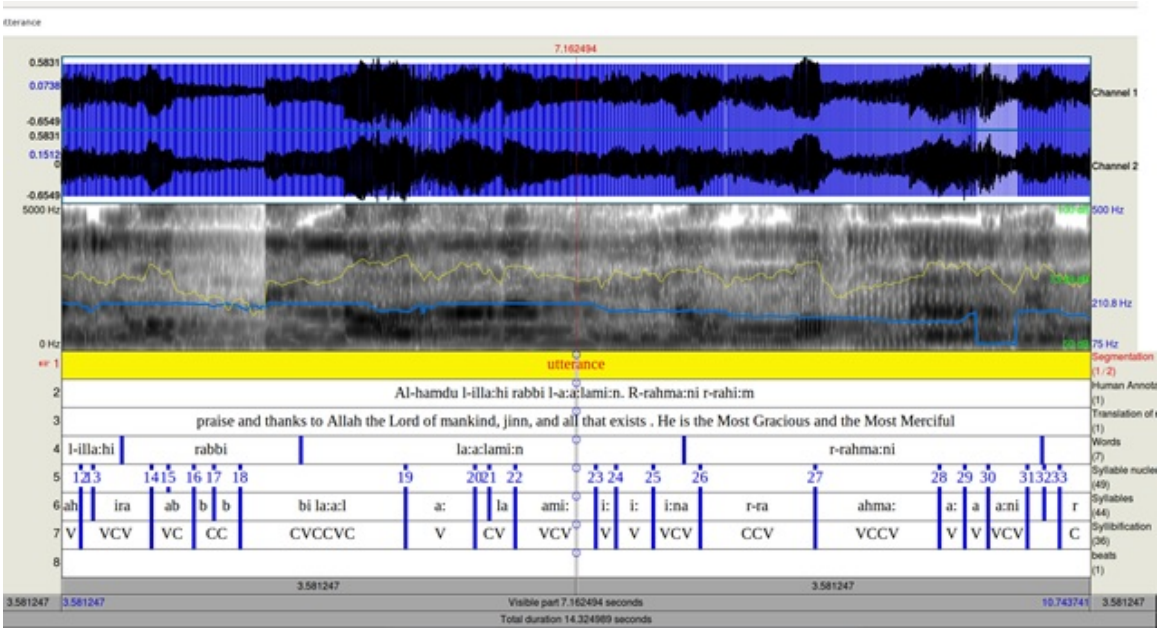

Figure 3: *The waveform and spectrogram of the sentence uttered by an Arabic native reciter Al-hamdu l-illa:hi rabbi l-a:a:lami:n. R-rahma:ni R-rahi:m annotated on seven tiers: 1) segmentation into utterance and silence segments (first top tier), 2) human transcription, 3) words 4) syllable nucleus, 5) syllables, 6) syllabification and 7) beats (bottom tier).*

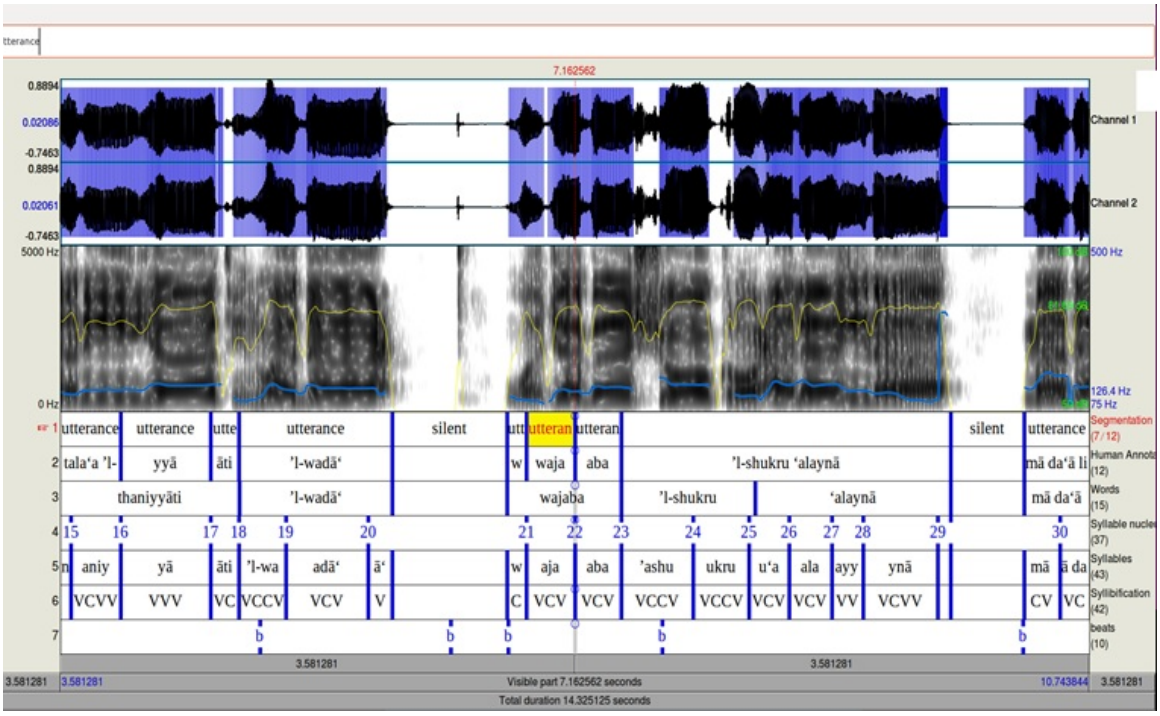

Figure 4: *The waveform and spectrogram of the sentence ala a 'l-badru alayn min thaniyyti 'l-wad wajaba 'l-shukru alayn m da li-l-lhi d a" annotated on seven tiers: 1) segmentation into utterance and silence segments (first top tier), 2) human transcription, 3) words 4) syllable nucleus, 5) syllables, 6) syllabification and 7) beats (bottom tier).*

signals (such as stress, pitch and intensity). Tracking these beats is so important for detecting the tempo of rhythm. To this end, beat locations of each utterance were estimated using a smooth in-

Figure 5: *Three tiers were shown for an utterance of "Al-hamdu l-illa:hi rabbi l-a:a:lami:n. R-rahma:ni R-rahi:m" which was uttered by Arabic Foreign Learner (AFL). The top tier presents the waveform, the middle tier shows the intensity and pitch contours, and the bottom tier includes the layers of features: segmentation, intensity peak features, intensity valleys, phrase boundaries and prominence.*

tensity contour of the filtered signal. Each beat is associated with the midpoint of each local rise in intensity contour (Cummins et al., 1998).

8 Automatic Rhythm Classification

Rhythm classification was done after determining speech rhythm features of native and non-native utterances. Such derived features were compute using syllable nucleus, beat locations, PPh and Prominence labels of speech corpus of Arabic religious and poem utterances read by AFL in different speaking styles. Extracted features of automatically identified labels of non-native Arabic speech were matched with that of Arabic native gold standard. After that Pearson correlation is computed to match between human scoring and automatic scoring. So, here in the classification task we use different rhythm features to measure to what extent the influence of such features on automatic scoring of rhythm proficiency of AFL. The classification task was done by scoring the degree to which the speech rhythm of learners utterance was similar to the Arabic native speech. stress and tone labels for AFL read speech in terms of their applicability for automated speech scoring. Both types of features have not been used in the context of automated scoring of non-native read speech to date. For this task Nave Bayes learning algorithm was used as a probabilistic linear classifier and a J48 classifier to compare the performance of the two classifiers. According to these scores, feedback reflecting the rhythm proficiency of each utterance could then be generated by the model of speech rhythm proficiency. Assessment the rhythm level of each utterance was achieved by comparing the generated scores of prediction and detection models and measuring the distance between the predicted rhythm and that which was detected. In the proposed study, a three-point scale ranging from (5) like-native, (4) semi-native to (3) poor was used to reflect the rhythmic proficiency level of learner speech using different speech contexts. The learners were encouraged through some form of feedback (positive or negative) to inform them of their rhythm errors that had to be corrected and improved.

9 Results

The results obtained from J48 and Nave Bayes classifiers are displayed in Table 2. The visualisation of the classification tree of J48 with the size of seven trees and four leaves is depicted in figure 6 . By comparing the results obtained from the J48 classifier which was trained with a full feature set of L2 utterances with those of the Nave Bays classifier we can observe a significant improvement. What can be clearly seen in table 2 is the growth of TP rate when using J48 for semi-native and poor classes. Similarly, F-measure rose when using J48 for all classes (Like-native, semi-native and poor).

Classifier	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
Nave Bays	1	0.5	0.3	1	0.5	.75	Like-Native
Nave Bays	0.677	0	1	0.677	0.8	1	Semi-native
Nave Bays	0	0	0	0	0	0.25	Poor
J48	0.833	0	1	0.833	0.909	0.962	Like-Native
J48	1	0.1	0.875	1	0.933	0.964	Semi-native
J48	1	0	1	1	1	1	Poor

Table 2: *Results from the J48 and Nave Bayes classifiers.*

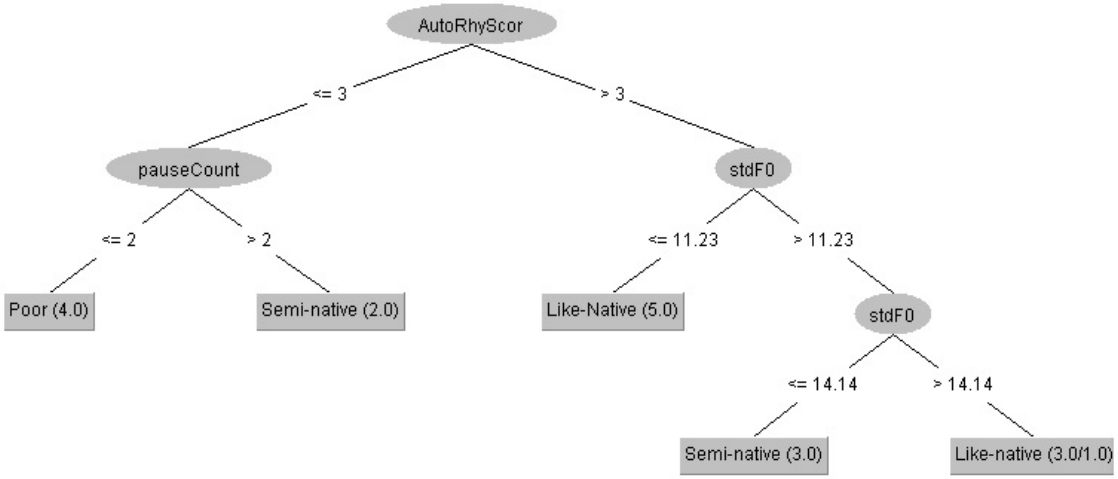

Figure 6: *The visualisation of classification tree of J48 classifier.*

More classifiers such as SVM and artificial neural network will be used in further work on the basis of these results.

10 Discussion, Conclusion and Future Work

We presented here an approach for classifying the rhythm of L2 utterances from different contexts performed with different speaking styles and rates. Results have shown that the J48 classifier was certainly more accurate and obtained results which were significantly improved compared to using Nave Bays. To improve classification accuracy, further improvement of this work will be done to extend the collected data in order to include more participants, including children and females from different languages such as Japanese and other stress-timed and syllable-timed languages. Moreover, other speaking context and styles will be covered in the next study to enrich the results. The results of this study are expected to be a starting point for improving SLaTE technologies. For example, acquisition skills to approximate stress-timed rhythm for L2 learners could be developed via establishing automatic religious and poem rhythm generation models for prosodic training. Furthermore, such models can be evaluated using other stress-timed languages such as English and Russian. In this paper we present the investigation of a number of rhythmic features computed using automatically labelled syllable nucleus, phrase boundaries, prominence and pitch information. Such features have not previously been used for the purpose of automatically scoring religious and poem read speech of AFL. Automatic scoring of these features were compared to scoring by an Arabic native expert and shows the deviation in the speakers rhythm from native rhythm of the target language. The average correlations of automatic proficiency scoring with human scoring was r=0.36. Overall, rhythmicity of joint melodic chanting of religious texts exhibit higher correlations, whereas the lowest correlation is observed for the rhythmicity of individual recitation of poem excerpts with (r=0.21). Particularly, features values of melodic chanting of joint recitation of native speakers had the highest correlation with those of non-native speakers with (r=0.97) while the correlation of automatic score of individual recitation styles of religious text with human raters was fairly low(r=0.41) due to the divergence of speech rates and influence of prosodic characteristics of L1 of non-native speakers. The results exhibit that values of pitch

and intensity contours of recitation style of religious verses achieved a high correlation of automatic rhythmicity score with the human score with(r=0.96). In this study, we used decision tree classifiers trained on two distinct styles (individual recitation and joint melodic chanting) of religious and poem contexts. However, the rhythm characteristics of oratory style of religious and politician contexts may present a high correlation score with human; consequently, we plan to train the classifiers with rhythmic features of such style and hope to improve the performance of rhythm prediction model to predict the most rhythmic style that is able to speed up the acquisition of native Arabic rhythm . Also,we plan to train SVM and Neural network classifiers with features of above mentioned rhythmic contexts and styles. Finding an average correlation with human proficiency scores will be our next step in terms of rhythm assessment of AFLs.

References

Muhammad Jamil Anwar, MM Awais, Shahid Masud, and Shafay Shamail. 2006. Automatic arabic speech segmentation system. *International Journal of Information Technology*, 12(6):102–111.

Juan Pablo Arias, Nestor Becerra Yoma, and Hiram Vivanco. 2010. Automatic intonation assessment for computer aided language learning. *Speech communication*, 52(3):254–267.

Jeesoo Bang, Sechun Kang, and Gary Geunbae Lee. 2013. An automatic feedback system for english speaking integrating pronunciation and prosody assessments. In *Speech and Language Technology in Education*.

Paul Boersma et al. 2002. Praat, a system for doing phonetics by computer. *Glot international*, 5.

Dmitry Bogdanov, Nicolas Wack, Emilia Gómez Gutiérrez, Sankalp Gulati, Perfecto Herrera Boyer, Oscar Mayor, Gerard Roma Trepat, Justin Salamon, José Ricardo Zapata González, and Xavier Serra. 2013. Essentia: An audio analysis library for music information retrieval. In *Britto A, Gouyon F, Dixon S, editors. 14th Conference of the International Society for Music Information Retrieval (ISMIR); 2013 Nov 4-8; Curitiba, Brazil.[place unknown]: ISMIR; 2013. p. 493-8.* International Society for Music Information Retrieval (ISMIR).

Lei Chen and Klaus Zechner. 2011. Applying rhythm features to automatically assess non-native speech. In *Twelfth Annual Conference of the International Speech Communication Association*.

Fred Cummins, Robert Port, et al. 1998. Rhythmic constraints on stress timing in english. *Journal of Phonetics*, 26(2):145–171.

Nivja H De Jong and Ton Wempe. 2009. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods*, 41(2):385–390.

Norberto Degara, Antonio Pena, Matthew EP Davies, and Mark D Plumbley. 2010. Note onset detection using rhythmic structure. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5526–5529. IEEE.

Monica Dominguez, Mireia Farrús, and Leo Wanner. 2016. An automatic prosody tagger for spontaneous speech. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 377–386.

Luciana Ferrer, Harry Bratt, Colleen Richey, Horacio Franco, Victor Abrash, and Kristin Precoda. 2014. Lexical stress classification for language learning using spectral and segmental features. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7704–7708. IEEE.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Debra M Hardison. 2004. Generalization of computer assisted prosody training: Quantitative and qualitative findings.

Tae-Yeoub Jang. 2009. Automatic assessment of non-native prosody using rhythm metrics: Focusing on korean speakers english pronunciation. In *Proc. of the 2nd International Conference on East Asian Linguistics*.

Pascale Lidji, Caroline Palmer, Isabelle Peretz, and Michele Morningstar. 2011. Listeners feel the beat: Entrainment to english and french speech rhythms. *Psychonomic bulletin & review*, 18(6):1035–1041.

Mostafa Ali Shahin, Julien Epps, and Beena Ahmed. 2016. Automatic classification of lexical stress in english and arabic languages using deep learning. In *INTERSPEECH*, pages 175–179.

Anna Sundström et al. 1998. Automatic prosody modification as a means for foreign language pronunciation training. In *Proc. ISCA Workshop on Speech Technology in Language Learning (STILL 98), Marholmen, Sweden*, pages 49–52.

Dávid Sztahó, Gábor Kiss, and Klára Vicsi. 2018. Computer based speech prosody teaching system. *Computer Speech & Language*, 50:126–140.

Joseph Tepperman and Shrikanth Narayanan. 2005. Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, volume 1, pages I–937. Citeseer.

Yi Xu. 2013. Prosodyproa tool for large-scale systematic prosody analysis. Laboratoire Parole et Langage, France.

Klaus Zechner, Xiaoming Xi, and Lei Chen. 2011. Evaluating prosodic features for automated scoring of non-native read speech. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 461–466. IEEE.