Rhythm Analysis of Second-Language Speech Through Low-Frequency Auditory Features

Jing Zheng[1*], *Joseph Tepperman*[2]

[1] University of Colorado, Boulder
[2] Rosetta Stone, USA

jing.zheng@colorado.edu, jtepperman@rosettastone.com

Abstract

Rhythm patterns play an important role in the perception of second-language (L2) speech. This paper presents a novel approach to evaluating L2 speech rhythm using low-frequency spectral features inspired by the rhythmogram auditory model. In this paper we investigate several new feature sets for use in training rhythm-centric acoustic models. By capturing information over suprasegmental linguistic units appropriate for rhythmic analysis (including syllables and prosodic feet), these novel features can outperform traditional features in detecting rhythm errors on the ISLE corpus of learner English by 5-15% absolute.

Index Terms: rhythm, language learning, stress, nonnative speech, auditory modeling, prosody

1. Introduction

Native-like rhythm patterns help make second-language (L2) speech intelligible. For example, the patterns of segment durations in English can indicate syllabic structure, lexical stress, and consonantal voicing distinctions, and hence can determine a word's identity [7]. Studies that have artificially imposed L2 rhythms onto native speech have found the intelligibility threshold of the speech in noise to degrade by about 4 dB [14] while upsetting native listener performance on speech identification tasks [18].

Though much headway has been made in automatic phoneme-level analysis of L2 pronunciation, including more than 10 years of automatic segmental scoring comparable to inter-listener agreement in perceived pronunciation quality [13], the field of linguistic rhythm has been stymied by the lack of an analysis framework for unifying the many diverse perceptual and auditory cues that characterize rhythmic patterns. We follow [23] in regarding rhythm as the perceived patterning of sound in short-term memory, i.e. with rhythmically salient events perceptually occurring at intervals between 100ms and 3s (the low-frequency range in which listeners perceive temporal events as rhythmic). This is in contrast to the theory that rhythmic typologies arise from intrinsic phonotactic differences among languages [3], an idea responsible for the many segment-duration-based rhythm metrics that dominate today's linguistic rhythm analysis literature, e.g. [15]. The reliability of these metrics and the validity of the theory behind them have been called into question increasingly often in recent publications, e.g. [1], though their main flaw may be the dependence on raw phonological durations in the absence of an auditory model.

In this work, we argue that some type of auditory model is a necessary part of the rhythm analysis task, in order to pinpoint rhythmically salient events from the acoustic signal itself. It is well-known that the centers of perceived syllabic "beats" in speech — the so-called p-centers — do not in general align with phonetic or syllabic boundaries, and they are dependent on many disparate factors such as amplitude envelope and fundamental frequency [26]. Moreover, traditional rhythm metrics based on segment durations (such as the PVI, %V, varcoC, etc.) are not able to detect L2 rhythm errors per se, but rather are designed to estimate gross differences in rhythmic patterns between populations or texts. Though some work has been done in template-based scoring of durational patterns [19], a model to capture the perceived rhythmic abstraction of speech acoustics is still to be desired. Here we propose a version of such a model, based partly on the rhythmogram system of analysis first proposed in [22]. This model relies on low-frequency acoustics alone and requires no linguistic knowledge of the signal. It also yields auditory features that are straightforward to incorporate into existing acoustic modeling frameworks such as Hidden Markov Models (HMMs). As we hope to demonstrate, these auditory features, designed specially to capture rhythmic information, are appropriate for modeling suprasegmental units, yielding improved detection of rhythm errors.

2. Low-frequency rhythmic information

Applying low-frequency features to speech recognition is hardly a new idea. The widely-used RASTA method of auditory-based speech processing [4] and its extension as the modulation spectrogram [6] have shown that pre-emphasizing low-frequency temporal structure in the speech signal can lead to significant gains in recognition accuracy under noisy conditions. Low-frequency modulation components have been found useful in other tasks as well, such as speaker verification [27] and music-speech discrimination [5]. The importance of low-frequency information can be explained in terms of the phonological characteristics encoded in the temporal motions of the larger speech production organs — the tongue, jaw, velum, and lips — which require roughly 50-300ms to reach the extreme ends of their movements [24]. From the perceptual side, these slow articulation rates overlap with the scale in which inter-event onsets are perceived as rhythmic [23].

Strangely, low-frequency auditory features have figured relatively insignificantly in the literature specifically devoted to speech rhythm, partly due to the dominance of duration-based rhythm metrics. There are exceptions: [25] used low-frequency Fourier analysis without an explicit auditory model to measure rhythms in conversational speech, and [16] derived rhythm features from the Hilbert-Huang Transform of low-frequency speech signals for emotional speech assessment.

* First author was an intern at Rosetta Stone at the time this work was completed.

But the prior low-frequency analysis model closest to the one proposed in this paper is that of the rhythmogram [22]. Described as an "auditory primal sketch," the rhythmogram is designed to detect the boundaries of perceptually salient rhythmic events via a full auditory model that includes an outer-middle ear filter, a filterbank representation of the cochlea, a model of inner-hair cell (IHC) firing in the auditory nerve, and a set of low-frequency modulation filters that approximates the multiresolution temporal response of the auditory cortex. Initially applied to impressionistic comparative analyses of speech, music, and poetry [21], more recent rhythmogram publications have used it to distinguish between hypothesized rhythm classes [9], to detect syllabic prominence [11], and to augment speech recognition lattice decoding [10]. For speed of computation in phrase-level analysis, [21] recommends replacing the cochlea and IHC stages with a rectified version of the raw speech signal, and then performing modulation filtering from there; this simplified approach has been followed in [11] and likewise will be adapted here. One aspect of the rhythmogram will be notably absent from this work: all prior rhythmogram-based publications have thrown away the rich spectral information in the modulation filter output in favor of detecting auditory "edges" in the signal and deriving features from those. Here we will use the full output of the auditory model as features for acoustic modeling.

3. Proposed auditory model and features

Our simple rhythmic auditory model (SRAM) is diagrammed in Figure 1. First the raw signal is rectified and passed through a first-order low-pass Butterworth filter with a passband edge of 50 Hz (to eliminate frequencies outside the range of rhythmic perception), and then resampled at 100 Hz (to reduce processing time). This decimated, low-frequency signal is intended to be a simplified approximation of summed IHC firing in the auditory nerve (hence the S in SRAM). It is next fed into a bank of 60 low-pass modulation filters of ever-widening passband, a model for multi-resolution processing in the auditory cortex. To temporally align the filter outputs relative to one another, these filters were all designed to have the same linear group delay (20 samples) using the least-squares IIR filter design algorithm proposed in [8]. Beginning at 0.3 Hz, the 60 filter passband edges were spaced by one semitone, with the widest passband edge falling at about 9 Hz, roughly covering the temporal range of rhythm perception from [23].

An example waveform and its SRAM output are displayed in Figure 2, alongside a conventional spectrogram over the full frequency range. The SRAM auditory filters are ordered with increasing passband edge from top to bottom. Rhythmically salient events (i.e. the p-centers of each syllable) are immediately visible (red = strongest). Essentially this is a graphical depiction of relative auditory prominence across the

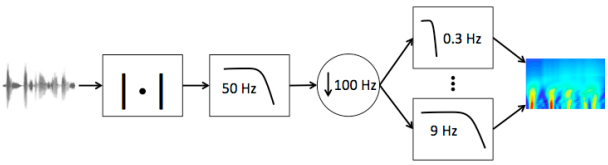

Figure 1: *Diagram of the proposed auditory model.*

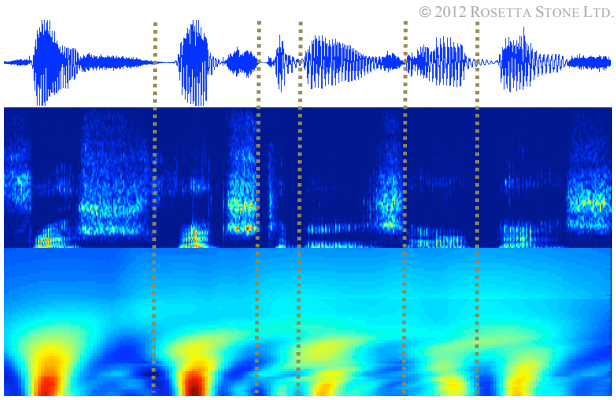

Figure 2: *The waveform (top) aligned with its spectrogram (middle) and the output of our proposed auditory model (bottom), for the phrase "Fresh vegetables are best." The dotted lines denote syllable boundaries. [Note: this is a color image]*

utterance on several simultaneous scales, as cued by low-frequency rhythmic information alone. Hierarchical phrasing is also suggested by the turquoise halos that appear to connect and organize the pattern of rhythmic events:

((fresh) ((vege- (ta- bles)) (are best)))

Though the relevant information is mostly found in the bottom half of this SRAM example, the entire range of filters is still necessary to capture the full spectrum of human rhythmic perception (as outlined in Section 2). Note that the perceptual centers do not necessarily align with syllabic onsets or nuclei (e.g. the center of "are" appears to be in the /ɹ/), so the SRAM is doing more than just finding the low-frequency "blobs" in the traditional spectrogram. Note also that the design requirement of identical phase distortion across the 60 modulation filters has removed the tendril-like forward-leaning tendency seen in prior studies' rhythmograms. Though [22] argues that this forward sweep indicates temporal masking effects in the short-term auditory store, it is more likely an artifact of the nonlinear group delay in [22]'s filterbank and can be problematic for waveform-feature alignment.

As input to conventional HMMs for abstracting speech acoustics, we now propose three sets of features derived from our auditory model. First is the raw output of the model itself, over the 60 modulation filters (SRAM), as shown in Figure 2. A direct extension would be to include delta and acceleration estimates on the frame level (SRAM_DA), to capture temporal dynamics. To reduce the size of this set, we tried Principal Component Analysis (PCA) on the SRAM_DA features, but this weakened the model's discriminatory power, suggesting that all 180 features were in fact useful. Due to the visible hierarchy of beat patterns in the SRAM output (see Figure 2), and inspired by work in HMM-based handwriting decoding, we propose extracting local gradient histogram features [17] from the SRAM image (SRAM_LGH) by treating time as horizontal pixels and the 60 modulation filters as vertical pixels. Using a left-right sliding window 18 pixels wide (i.e. 18 samples in the 100 Hz resampled signal) with a 1-pixel overlap, we split each window into a 3x3 grid of cells and sum the pixel-level gradient magnitudes over 6 angular bins, leading to 3x3x6=54 features per window (this is lower resolution than the 128 features proposed in [17], since our shapes are generally simpler than handwriting). In a sense, the

LGH features are intended to capture the multiscale rhythmic "handwriting" of speech, including its temporal dynamics. When used in HMMs (which ordinarily lack an explicit duration model), all of these feature sets are intended to encode the prominence and phrasing patterns that characterize iconic rhythms; implied timing is secondary. These novel feature sets are summarized in Table 1. As a baseline, we also calculated MFCC and PLP sets — traditional auditory features that are not rhythm-specific. These included energy, delta, and acceleration coefficients, in 39 dimensions. Since the SRAM features are resampled to 100 Hz, the proposed and baseline sets are not defined on the same scale. Fusing them into a larger feature set is nontrivial, and we will reserve that for future work.

4. Rhythm error detection experiments

To compare the proposed feature sets and evaluate their rhythm analysis power relative to the baselines, we now report the results of experiments in detecting errors in lexical stress patterns — one class of rhythm errors believed to be a source of negative L1 transfer effects [12].

4.1. Corpus

The Interactive Spoken Language Education (ISLE) corpus [2] is designed for developers of English language learning systems. It comprises 18 hours of read speech from 46 intermediate learners of British English, split evenly between native Italians and Germans. Many of the utterance prompts were designed to accentuate minimal-pair stress contrasts (e.g. "Children often rebel against their parents"), leading to some common L2 rhythm mistakes. We used only Blocks D, E, F, and G, which come automatically aligned and manually transcribed for phoneme and stress errors.

4.2. Method

Detecting errors in L2 lexical stress patterns requires training models for stress categories on the syllable level, or for categorical stress patterns above the syllable level. Using the features described in Section 3, we trained the following sets of suprasegmental HMMs:

- STR: a simple set of binary stress models to describe any syllable (i.e. either **str** or **unstr**)
- STR-V: as in STR, but with each syllable's nuclear vowel context (e.g. "having" = **str-**/æ/ , **unstr-**/ɪ/)
- FOOT: stress pattern across the syllables of a prosodic foot (e.g. a three-syllable left-headed foot would map to the model **str-unstr-unstr**)

Note that fully context-dependent syllable models were not trained due to sparsity in the ISLE corpus.

Table 1. *Proposed auditory features. See Section 3 for details.*

feature set	***description***	***dimensionality***
SRAM	output of model	60
SRAM_DA	…with first and second derivatives	180
SRAM_LGH	local gradient histogram of model	54

Table 2. *EER (%) over various model-feature combinations.*

	model		
feature set	STR	STR-V	FOOT
SRAM	47.74	45.02	39.74
SRAM_DA	**45.02**	42.53	37.17
SRAM_LGH	45.57	45.02	**34.99**
MFCC	52.49	37.48	67.50
PLP	50.08	**32.81**	50.31

HMM training used a standard flat-start initialization and iterative reestimation procedure. Since all monosyllabic words in the ISLE corpus come erroneously transcribed as stressed in a phrase context, we designed the model reestimation so that any function word could be re-labeled as unstressed, if that improved the likelihood of the data given the model. Acoustic silence and garbage models were also trained for each set. Because the syllabic length of a foot can vary considerably, the number of states in each foot-level model was determined by assigning 5 states to the first syllable and 4 states to any additional syllables. All other HMMs used a standard 5-state left-right topology, with the number of Gaussian mixtures per state varying from 2 to 256 according to the number of training instances available. The train set consisted of 36 speakers, and the test set had the remaining 10 (both split evenly between Italians and Germans). To isolate the effects of rhythm errors, we removed test utterances with segmental mispronunciations.

To target rhythm variations over phrase-level patterns, stress error detection was performed on the level of the sentence prompt. A rhythm score was estimated as a standard likelihood ratio for each sentence, where the numerator was the likelihood of the target prompt, and the denominator was a decoding loop over all HMMs in the given model set. For the FOOT models, the denominator loop was actually a sequence of loops, one for each foot in the target: each loop would decode only feet the same length as, or one syllable shorter than, the corresponding foot in the target prompt sequence. A higher rhythm score would indicate the learner's stress pattern was closer to that expected by the transcribers and demanded by the text. Error detection was done by comparing these likelihood ratios to a varying global threshold. The equal error rates (EER) are given in Table 2, over all combinations of feature sets and models.

4.3. Discussion

The goal of these experiments was to compare the novel rhythm-based feature sets to some baseline auditory features (MFCCs and PLPs), over various model time scales and contexts. In Table 2, we see that the STR-V set, the only one that uses phoneme-level context, is also the only one where the MFCC and PLP features outperform the three SRAM sets. For STR and FOOT models, which are based on stress baseforms alone, the novel features perform better than either MFCC or PLP (these improvements are significant on the 95% level using a one-tailed t-test). The temporal derivatives in the SRAM_DA and SRAM_LGH sets improve over SRAM alone, but SRAM_LGH requires considerably fewer features per frame.

The interpretation of Table 2 seems clear: the SRAM is more appropriate for modeling stress patterns over suprasegmental levels — syllables and feet — while MFCCs and PLPs can better quantify the short-time auditory spectra that characterize phonemes. This explains the huge improvement in MFCC and PLP performance with the addition of vowel context as part of the model (i.e. between STR and STR-V). The baseline spectral features may also be better at capturing subtle differences in vowel quality that accompany syllabic stress shifts in rhythm errors. But the suprasegmental aptitude of the novel features is most clearly seen in the improvement in the SRAM performance when the linguistic unit grows to the FOOT level (differences between FOOT and STR models for the SRAM sets were significant on the 95% level). The performance difference between the two best setups — PLP/STR-V and SRAM_LGH/FOOT — was not statistically significant, indicating that the baselines and SRAMs have comparable performance when paired with the appropriate model. As these are sentence-level results, how they would compare to state-of-the-art performance on the syllable level is unclear, though there is still room for other features such as word-level context and explicit pitch and duration features [11],[20].

5. Conclusion

While MFCCs and PLPs are the auditory features of choice for phoneme-level acoustic modeling, low-frequency spectral information can be more appropriate for the suprasegmental scales of speech rhythm patterns. In the absence of segment-level contextual info, our proposed auditory features outperform the baselines in verifying sentence-level L2 rhythms by 5-15% absolute.

SRAM features are far simpler to compute than MFCCs or PLPs. Unlike duration-based rhythm metrics, SRAMs offer a perception-based analysis of rhythm, they naturally lend themselves to models of linguistic abstraction such as HMMs, and they can be used to detect L2 rhythm errors. And unlike any of the feature sets mentioned here, they can pinpoint the perceptual centers of syllabic beats (recall Figure 1), for a richer rhythmic analysis of which this paper is only the beginning. Future work can extend this model by introducing other perceptual cues (e.g. pitch and duration, phonological and syntactic information, etc.) and by using SRAM features to "parse" perceived rhythmic groupings, as hinted by the visible hierarchy in Figure 1.

6. References

[1] A. Arvaniti, "Rhythm, timing, and the timing of rhythm," *Phonetica*, 66:46-63, 2009.

[2] E. Atwell, P. Howarth, and C. Souter, "The ISLE corpus: Italian and German spoken learner's English," *ICAME Journal*, 27:5-18, 2003.

[3] R.M. Dauer, "Stress-timing and Syllable-timing Reanalyzed," *Journal of Phonetics*, 11:51-62, 1983.

[4] H. Hermansky and N. Morgan, "RASTA Processing of Speech," *IEEE Transactions on Speech and Audio Processing*, 2(4):578-589, 1994.

[5] S. Karnebäck, "Discrimination between Speech and Music based on a Low Frequency Modulation Feature," *European Conference on Speech Communication and Technology*, 2001.

[6] B.E.D. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Communication*, 25:117-132, 1998.

[7] P. Ladefoged, *A Course in Phonetics*, Thomson, Boston, 2006.

[8] M. Lang, "Algorithms for the Constrained Design of Digital Filters with Arbitrary Magnitude and Phase Responses," Österr, Kunst-und Kulturverl., 2000.

[9] C.S. Lee and N.P.M. Todd, "Towards an Auditory Account of Speech Rhythm: Application of a Model of the Auditory 'Primal Sketch' to Two Multi-language Corpora," *Cognition*, 93(3):225-254, 2004.

[10] B. Ludusan, S. Ziegler, and G. Gravier, "Integrating Stress Information in Large Vocabulary Continuous Speech Recognition," *Proc. of Interspeech*, Portland, 2012.

[11] B. Ludusan, A. Origlia, F. Cutugno, "On the Use of the Rhythmogram for Automatic Syllabic Prominence Detection," *Proc. of Interspeech*, Florence, 2011.

[12] H. Meng, C.-Y. Tseng, M. Kondo, A. Harrison, and T. Viscelgia, "Studying L2 Suprasegmental Features in Asian Englishes: A Position Paper," *Proc. of Interspeech*, 2009.

[13] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic Scoring of Pronunciation Quality," *Speech Communication*, 30(2):83-93, 2000.

[14] H. Quené and L.E. van Delft, "Non-native durational patterns decrease speech intelligibility," *Speech Comm.*, 52:911-918, 2010.

[15] F. Ramus, M. Nespor, and J. Mehler, "Correlates of linguistic rhythm in the speech signal," *Cognition*, 73:265-292, 1999.

[16] F. Ringeval and M. Chetouani. "Hilbert-Huang Transform for Non-Linear Characterization of Speech Rhythm," *NOLISP*, 2009.

[17] J.A. Rodríguez-Serrano and F. Perronnin, "Handwritten word-spotting using hidden Markov models and universal vocabularies," *Pattern Recognition*, 42:2106-2116, 2009.

[18] K. Tajima, R. Port, and J. Dalby, "Effects of temporal correction on intelligibility of foreign-accented English," *Journal of Phonetics*, 25:1-24, 1997.

[19] J. Tepperman, T. Stanley, K. Hacioglu, and B. Pellom, "Testing Suprasegmental English Through Parroting," *Proc. of Speech Prosody,* Chicago, 2010.

[20] J. Tepperman and S. Narayanan, "Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners," *Proc. of ICASSP,* Philadelphia, 2005.

[21] N.P.M. Todd and G.J. Brown, "Visualization of Rhythm, Time and Metre," *Artificial Intelligence Review*, 10:253-273, 1996.

[22] N.P.M. Todd, "The Auditory 'Primal Sketch': A Multiscale Model of Rhythmic Grouping," *Journal of New Music Research*, 23(1):25-70, 1994.

[23] W.A. Sethares, *Rhythm and Transforms*, Springer, 2007.

[24] K.N. Stevens, *Acoustic Phonetics*, MIT Press, 2000.

[25] S. Tilsen and J. Keith, "Low-frequency Fourier analysis of speech rhythm," *JASA*, 124(2):34-39, 2008.

[26] R. Villing, "Hearing the Moment: Measures and Models of the Perceptual Centre," PhD Dissertation, NUI Maynooth, 2010.

[27] S. van Vuuren and H. Hermansky, "On the importance of components of the modulation spectrum for speaker verification," *Proc. of ICSLP*, Sydney, Nov. 1998.