SPEECH RHYTHM GUIDED SYLLABLE NUCLEI DETECTION

Yaodong Zhang and James R. Glass

MIT Computer Science and Artificial Intelligence Laboratory
Cambridge, Massachusetts 02139, USA
{ydzhang,glass}@csail.mit.edu

ABSTRACT

In this paper, we present a novel speech-rhythm-guided syllable-nuclei location detection algorithm. As a departure from conventional methods, we introduce an instantaneous speech rhythm estimator to predict possible regions where syllable nuclei can appear. Within a possible region, a simple slope based peak counting algorithm is used to get the exact location of each syllable nucleus. We verify the correctness of our method by investigating the syllable nuclei interval distribution in TIMIT dataset, and evaluate the performance by comparing with a state-of-the-art syllable nuclei based speech rate detection approach.

Index Terms— speech rhythm, syllable nuclei detection

1. INTRODUCTION

Human speech has been observed by linguists as having a temporal rhythm that can be characterized by placing a perceptual "beat" around successive syllables [1]. This periodic "speech rhythm" provides a syllable-level characterization of the speech signal that approximates the number of segmental elements in the utterance [2, 3]. Most speech processing methods ignore this source of information however, and almost exclusively use short-time spectral analysis of the speech signal. Since relying on a single source of information can lead to brittle behavior in conditions of uncertainty, our objective was to find a way to measure speech rhythm, and to incorporate it into a speech processing task.

Our initial investigations have focused on the problem of syllable nuclei detection. Although most speech recognition strategies employ frame-based analysis, a parallel syllable-level analysis stream can provide an additional source of constraint [4]. Syllabic nuclei are arguably one of the more reliable cues to be found in noisy speech, and as such can provide reliable anchor points for subsequent analysis [5, 6]. By counting the number of syllable nuclei in an utterance, we can get a measure of the speaking rate, which can be useful for characterizing speaking style, selecting appropriate acoustic models, or speaker adaptation.

Most existing methods for detecting syllable nuclei measure short-time acoustic features of the speech signal; produce a continuous curve of extracted features; and perform peak and/or valley detection to determine the numbers of possible vowels or syllable onsets [7]. Due to phonological effects or environmental noise, finding syllabic peaks in any feature representation is inherently error-prone. For example, spurious peaks can be caused either by background noise or by non-vocalic sounds; neighboring syllabic peaks can fuse because they are too close or the windowing parameter is incorrectly set. Although solutions to some of these problems have been proposed [4, 8], the resulting algorithms can be very sensitive to parameter settings, and correspondingly less robust on new data.

In this work we attempt to estimate the instantaneous speech rhythm and incorporate this information into a syllabic nuclei detection algorithm. At the time we perform peak detection on a feature envelope, we also estimate the instantaneous speech rhythm. By assuming that a normal speaker will not dramatically change their speech rhythm within an utterance (an assumption we discuss later), we use the current speech rhythm to predict the next possible time period where a syllable nucleus might appear. Within this period, we perform parameter-free peak detection to locate syllabic nuclei. We investigate our approach on the TIMIT dataset by measuring interval distributions of upcoming syllables both with and without speech-rhythm scaled durations. We also develop a robust syllabic nuclei detection method; report its performance on TIMIT; and compare results to a state-of-the-art method. Our experimental results show that the rhythm-based method significantly improves syllabic nuclei detection performance, and shows promise for other forms of speech processing.

In the following section, we will give a detailed description of our syllable nuclei detection algorithm. Experimental results will be reported in Section 3. We conclude and discuss ideas for future work in Section 4.

2. ALGORITHM DESIGN

Our speech-rhythm guided syllable nuclei detection algorithm can be divided into two main stages. In the first stage, similar to conventional methods, we apply envelope analysis on the input speech signal. In the second stage, we estimate the speech rhythm on the signal envelope to help us with syllable nuclei detection. A flowchart of our proposed algorithm in shown in Figure 1. Each of the stages is described in more detail in the following three subsections.

2.1. Envelope Analysis

To filter out noise at different frequency levels, the peripheral auditory based band-pass filter has been widely used in analyzing speech signals. In our approach, we use an equivalent rectangular bandwidth filter (ERB) to first split the wave signal into 20 channels. The ERB filter can be viewed as a series of gammatone filters operating on the ERB rate scale. The output signal of the i-th ERB filter is

$$x_i(t) = t^{m-1} \exp(-2\pi b_i t) \cos(2\pi CF_i^{nor}) H(t) \qquad (1)$$

where m is the filter order (in our implementation, we set $m = 4$), b_i is the bandwidth, CF_i^{nor} is the center frequency converted from

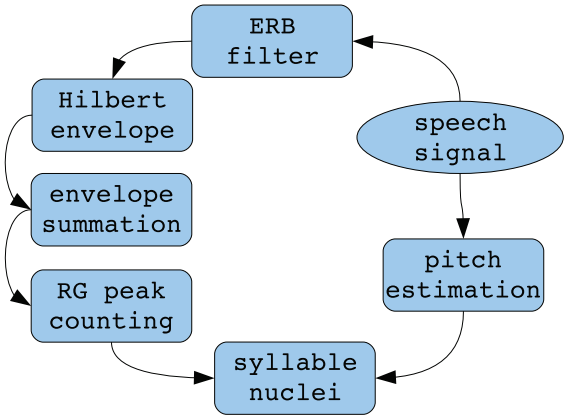

Fig. 1. Algorithm flowchart (RG stands for rhythm guided).

the ERB scale, and $H(t)$ is the unit step signal. For a more detailed description of the ERB filter, we refer readers to [9].

After band-pass filtering, we extract the envelope $E_i(t)$ of each channel. Let $x_i(t)$ be the signal in the i-th channel. To compute the envelope, we apply the Hilbert transform on $x_i(t)$ to get the magnitude $X_i(t)$ as

$$X_i(t) = x_i(t) + i \cdot \mathcal{H}(x_i(t)) \tag{2}$$

where $\mathcal{H}(\cdot)$ denotes the Hilbert transform. The envelope of the i-th channel can be calculated by $E_i(t) = |X_i(t)|$. In order to reinforce the energy agreement of each channel, we first normalize each channel envelope and then sum them to get the total envelope representation $E(t)$ of the input speech signal $E(t) = \sum_{i=1}^{N} E_i(t)$ where N is the number of channels.

Unlike some previous methods, we do not choose to do sub-band selection, and we use direct summation instead of sub-band temporal correlation of the energy envelopes. There are two main reasons for this. First, since we are using a two-stage method, the more sub-bands we use to contribute the total envelope, the more information we can use for speech-rhythm estimation. Second, we noticed that with temporal correlation, energy peaks may occur a little bit later than their actual location. The amount of delay is dependent on the length of the correlation window. Although this delay does not affect the total number of peaks, it does affect the accuracy of our subsequent speech-rhythm estimate. Therefore, we chose not to perform correlation so as to not interfere with the second stage processing.

2.2. Rhythm Guided Peak Counting

A perfect peak counting algorithm would find all true positive peaks and make no false predictions. In most peak picking methods, there are two important parameters that determine peak selection [10]. The first is the ratio of the height of a peak over the highest peak in the current working set. The other is the ratio of the contrast of a peak with its immediately neighboring valleys. The first parameter uses global information while the second parameter may vary on a case by case basis, even within an utterance. In addition, due to the various sources of noise, it can be difficult to find parameter thresholds that avoid all spurious peaks while detecting all correct peaks. Thus, we seek to use the speech rhythm information to avoid requiring these two parameters in processing the spectral envelope. The basic procedure is to use a conventional method to find the first two syllabic nuclei via envelope peaks; then estimate the instantaneous speech rhythm based on these two nuclei; and then subsequently predict intervals where the next syllable nucleus may appear; finally, we use a simple slope based peak detection algorithm in each interval that avoids the use of any parameters or thresholds. This simple peak detection and speech rhythm estimation are performed repeatedly until the end of the utterance is reached.

To clearly and efficiently represent the speech rhythm, we turn the speech rhythm estimation into a sinusoid function fitting problem. Given a peak location set P, we fit a sinusoid function F_{k_1,k_2}, with frequency, k_1, and phase offset, k_2, of which peak locations are matched to the peak locations in P. The target sinusoid function is

$$F_{k_1,k_2}(x) = sin(k_1 \cdot x + 2\pi \cdot k_2) \tag{3}$$

By using a least mean squares fitting scheme, the objective function can be written as

$$\{k_1, k_2\} = \arg\min_{k_1,k_2} \frac{1}{|P|} \sum_{i=1}^{|P|} (1 - F_{k_1,k_2}(p_i))^2 \tag{4}$$

where $p_i \in P$ denotes the location of peak p_i and $k_2 \in [0, 1)$. Using these notations, a stepwise description of the syllable nuclei detection method is as follows:

- **Step 1** After the i-th peak p_i is found, add p_i into P
- **Step 2** Based on the observed peaks in P, perform least mean squares fitting on P to get the current best k_1^i and k_2^i
- **Step 3** Align all the p_i in P to the nearest x_j and find x_{p_i} representing the sinusoid peak location to which the newly found p_i is assigned
- **Step 4** Calculate the smallest x_s where $F_{k_1^i,k_2^i}(x_s) = 1$ and $x_s > x_{p_i}$
- **Step 5** Run slope based peak counting within the range $[x_{p_i}, x_s + \frac{3\pi}{k_1^i}]$. If we have multiple peak candidates, pick the maximum one. Thus, we only allow one peak in this range.
- **Step 6** If a peak is found, go to Step 1. If not, set $x_{p_i} = x_s$ and go to Step 5. Repeat until reaching the end of utterance.

Note that we need at least two peaks to estimate the first set of k_1 and k_2. We initialize k_1 to be $\frac{2\pi}{k_1} = 100ms$ to avoid the uninteresting solution of large k_1. We tried both the standard contrast based and simple slope based peak counting algorithm and found that the selection of these two algorithms has little effect on the results, especially when we consider the overall precision and recall. We illustrate the first three steps in our algorithm on a TIMIT utterance in Figure 2. The blue curve is the extracted Hilbert envelope; the red circles consist of the current peak location set P. The sinusoid function, shown in magenta, is the fitted estimate of the current speech rhythm. The black dotted lines correspond to vowel boundaries in the phonetic transcription. From top to bottom, the figure shows the first three iterations of our algorithm. In the top plot, the first two peaks corresponding to the vowels /eh/ and /ax/ have been found without rhythm. From their locations, a rhythm sinusoid is estimated and extended into the future. In the middle plot, the next peak has been identified in the /ao/ vowel, which corresponds to the maximum peak in the interval under location (corresponding to 1.5 cycles of the rhythm sinusoid from the last peak). The rhythm sinusoid is re-estimated and extended into the future. In the bottom plot, the next peak has been located in the /ih/ vowel. This particular peak could have been easily missed due to its small size.

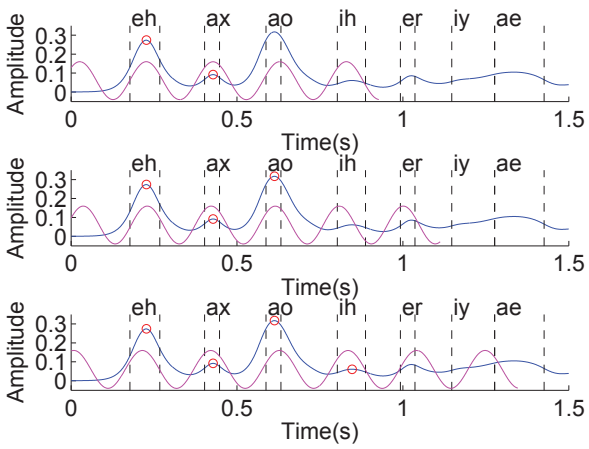

Fig. 2. Example of speech-rhythm based detection of syllable nuclei. See text for details.

2.3. Pitch Verification

We observe that it is possible for our rhythm guided approach to find spurious peaks at the beginning or end of an utterance, or it can place some peaks in utterance internal pauses. Thus, pitch verification is used to remove spurious peaks in unvoiced regions. For both methods after all peaks are detected, a removal operation is performed if we find a peak being located in a highly likely unvoiced region.

3. EVALUATION

In order to evaluate the speech rhythm-based method we designed two sets of experiments to 1) verify the correctness of our rhythm estimation idea; 2) compare the syllable nuclei detection performance with state of art techniques. The TIMIT dataset, which includes a total of 6300 phonetically-transcribed read-speech utterances, is used in both sets of experiments. Since the basic TIMIT corpus has no syllable segment annotation, we only consider vowels as syllable nuclei in all experiments.

3.1. Syllable-Nuclei Intervals

In order to establish the merit of the speech rhythm idea we examined the durations between nearby vowels in the TIMIT corpus, where vowels represented syllable nuclei. We gathered statistics on all syllable pair and triple sequences, measuring the interval between the first and last vowel center of each sequence. As shown in the left plot of Figure 3, we are thus effectively measuring the syllable nuclei intervals (SNIs) of adjacent syllables (shown in blue), and of those separated by exactly one syllable (shown in red). Note that from any given vowel center, the expected interval to the next vowel center is approximately 200ms, and an additional 200ms to the following vowel center. The plot clearly shows tremendous overlap in the two distributions however, so that there is a tremendous range where either one or two syllabic nuclei could occur. This observation explains why previous approaches to syllabic nuclei detection often resorted to elaborate peak picking selection methods to decide where peaks could occur.

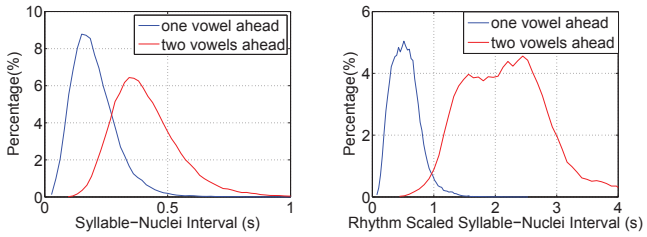

Fig. 3. Distribution of Syllable-Nuclei Intervals in TIMIT and their corresponding rhythm-scaled versions.

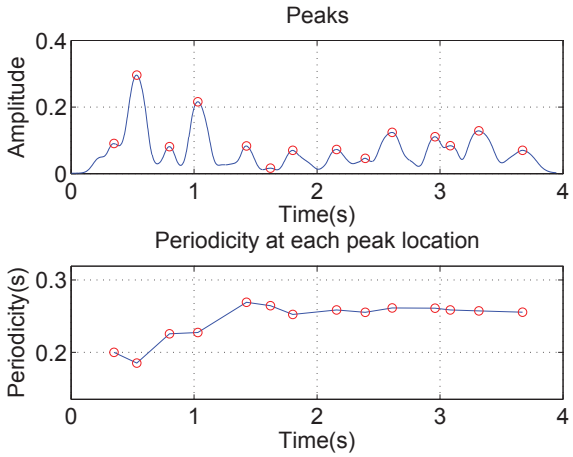

Fig. 4. Example of estimated instantaneous rhythm periodicity for a single TIMIT utterance. See text for details.

In an ideal case, if we knew the regular speech rhythm of syllabic nuclei, we would be able to better predict where the next syllable nucleus would occur. We can approximate this rhythm concept with our computed sinusoid frequency k_1 and use it to normalize the SNIs that were measured previously. Specifically, we scale each interval by the utterance-specific factor of $3\pi/k_1$, resulting in a dimensionless quantity which is plotted on the right side of Figure 3. This plot shows that the majority of SNIs for immediately adjacent syllables occur within an interval of $3\pi/k_1$ of an existing syllabic nucleus (blue). It also shows much less overlap with SNIs of following syllables (red). This result motivated our approach for syllabic nuclei detection, allowing us to avoid thresholds or parameters. The only parameter we selected was the value of $3\pi/k_i$.

The rhythm estimates used in Figure 3 were estimated over an entire utterance. A plot computed with rhythm estimates computed iteratively shows very similar results. We have also found that the estimates for rhythm converge relatively quickly. Figure 4 shows how the rhythm sinusoid periodicity ($2\pi/k_i$), changes over time as it is computed in a left-to-right fashion over an utterance. With each new peak detection (apart from the first two), shown in the upper plot of the figure, the sinusoid is re-estimated, and the period is plotted in the lower part of the figure. After several peaks have been located, the rhythm estimate becomes fairly stable, although the detection region still allows the detection of closely spaced syllable nuclei. Note that the default starting periodicity is 200ms.

Table 1. Syllable Nuclei Detection Comparison on TIMIT

	TCSSC	nRG	RG
Best Recall	0.8606	0.7997	**0.8659**
Best Precision	0.9969	**0.9984**	0.9886
Best F-measure	0.9021	0.8858	**0.9207**

3.2. Performance Comparison

To demonstrate the performance of our rhythm-guided (RG) syllable nuclei detection approach, we compared it against the state-of-the-art syllable nuclei based speaking rate estimation method TCSSC (temporal correlation and selected sub-band correlation) [8]. According to a recent comparative study paper of speaking rate [7], TCSSC has the best performance out of 8 different systems. In addition, since the RG method and TCSSC have different signal processing modules, we built another system (nRG) that applies the conventional contrast based peak counting algorithm to the Hilbert envelope data without any rhythm guiding. Our intent was to quantify the degree to which rhythm information can help with syllable nuclei detection. We used the same ESPS pitch tracker [11] for both methods.

The results of our vowel detection experiments are reported in Table 1 in terms of the best overall recall, precision, and F-measure that could be achieved with each approach. A successful detection meant that there was exactly one syllabic peak detected within a 50ms window of an actual vowel.

Since the TCSSC and nRG method used conventional peak picking, we tested a variety of different parameters to optimize performance. The only nRG parameter that was used was the $3\pi/k_1$ search interval, which was held fixed. All three methods had signal processing parameters that were varied to optimize performance.

The results indicate that in the best case scenario the three methods can locate between 80-87% of the vowels, and in another best case scenario can locate vowels with very high precision of 99%. The TCSSC and RG methods can achieve almost the same best recall performance, although the RG method requires significantly less parameter tuning.

Adding rhythm-based information clearly helps with recall and overall F-measure, although it seems to reduce best case precision over the nRG method. Overall, the best case F-measure showed the RG method outperformed both TCSSC and nRG methods. Given that TCSSC produced better recall results than nRG, it will be interesting to explore hybrid methods that combine elements of both the TCSSC and RG methods.

4. SUMMARY AND FUTURE WORK

This paper has presented our initial efforts at extracting and using speech rhythm information for the task of syllable nuclei detection. We presented a method to efficiently estimate the instantaneous speech rhythm, and to use it to help with the subsequent syllable nuclei detection. We demonstrated the potential usefulness of the instantaneous speech rhythm frequency by normalizing syllable nuclei intervals from vowels in the TIMIT corpus. We then performed a comparison between our method, both with and without speech rhythm information, compared to the current best-reported syllable nuclei detection approach, and showed that the rhythm-guided improved overall performance, while essentially eliminating parameters that need to be tuned to a corpus.

While the results we have observed are encouraging, they are preliminary, and there is much room for further improvement. For example, it is possible that a signal processing module that enhances syllabic peaks could improve the performance of the rhythm-guided syllable detection. More importantly however, we have begun to examine spontaneous speech, or long audio segments. It is clear that without proper compensation for natural pauses, the speech rhythm estimate will be gradually biased to a slower rhythm than is actually the case. Using the rhythm estimator only the presence of speech (i.e., with a voice activity detector [12]), should ameliorate this situation. Finally, we are interesting in developing a probabilistic framework for thinking about speech rhythm, and how it may be incorporated into other speech processing applications.

Since there is nothing inherently language specific about the speech-rhythm method, we have begun to analyze a variety of different languages, and have been encouraged by the initial results. Ultimately, we believe that speech-rhythm information should be helpful for speech recognition as well. We are beginning to explore its utility in landmark-based speech recognition [13], within the context of language independent processing.

5. REFERENCES

[1] R. F. Port, "Meter and speech," *Journal of Phonetics*, vol. 31, pp. 10–16, 2003.

[2] M. C. Brady and R. F. Port, "Quantifying vowel onset periodicity in japanese," in *Proc. International Conference on Phonetic Science*, 2007.

[3] K. Tajima, A. Zawaydeh, and M. Kitahara, "A comparative study of speech rhythm in arabic, english and japanese," in *Proc. International Congress of Phoneic Sciences*, 1999.

[4] N. Morgan and E. Fosler-Lussier, "Combining multiple estimators of speaking rate," in *Proc. ICASSP*, 1998, pp. 729–732.

[5] S. Wu, M. L. Shire, S. Greenberg, and N. Morgan, "Integrating syllable boundary information into speech recognition," in *Proc. ICASSP*, 1997, pp. 987–990.

[6] C. D. Bartels and J.A. Bilmes, "Use of syllable nuclei locations to improve asr," in *Proc. ASRU*, 2007, pp. 335–340.

[7] T. Dekens, M. Demol, W. Verhelst, and P. Verhoeve, "A comparative study of speech rate estimation techniques," in *Proc. INTERSPEECH*, 2007, pp. 510–513.

[8] D. Wang and S.S. Narayanan, "Robust speech rate estimation for spontaneous speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2190–2201, 2007.

[9] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, Academic, 1997.

[10] O. Lartillot and P. Toiviainen, "Mir in matlab: A toolbox for musical feature extraction from audio," in *Proc. ICMIR*, 2007, pp. 22–27.

[11] D. Talkin, "A robust algorithm for pitch tracking (rapt)," in *Proc. ICASSP*, 1983, pp. 1352–1355.

[12] J. Ramirez, J. C. Segura, J. M. Gorriz, and L. Garcia, "Improved voice activity detection using contextual multiple hypothesis testing for robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2177–2189, 2007.

[13] J. R. Glass, "A probabilistic framework for segment-based speech recognition," *Computer Speech and Language*, vol. 17, pp. 137–152, 2003.