Robust Acoustic-Based Syllable Detection

Zhimin Xie, Partha Niyogi

Department of Computer Science
University of Chicago, Chicago, IL
zxie@cs.uchicago.edu, niyogi@cs.uchicago.edu

Abstract

In this paper, we describe a method to detect syllabic nuclei in continuous speech. It employs two basic and robust acoustic features, periodicity and energy, to detect syllable landmarks. This method is evaluated on TIMIT, noise additive TIMIT and NTIMIT datasets with typical total error rates of around 30% in all the datasets, except for extremely adverse 0dB signal-noise-ratio environments, while HMM-based systems degrade rigorously. Based on the landmarks, a vowel classifier is further constructed and achieves the same performance as HMM-based systems.

Index Terms: syllable detection, robustness, vowel classification.

1. Introduction

Motivation: We consider the problem of detecting syllabic nuclei directly from the speech signal in absence of any higher level (i.e. word level or syntactic level) linguistic cues. We are motivated by three considerations. First, researchers in language acquisition have considered the problem of how children might segment the phonological stream into word boundaries and learn the words of their native language. It appears that the statistics of transitions between syllabic units are employed by infants as young as 8 months [1]. If one wishes to have a computational account of the mechanisms by which this is achieved, one will need to describe the first step of extracting syllabic units from the signal. Our paper presents results in that direction. Second, in phonology, the syllable has had a long tradition of inquiry associated with it. Notions such as sonority hierarchies have been developed to describe the patterning of syllabic nuclei in a phonological stream. Quantities like stress, tone, and prosody live at the syllabic tier of the phonological representation. Our work may be regarded as an investigation of acoustic correlates of syllabic contours to tie the phonological notions of the syllable to acoustic and phonetic properties of the speech signal. Finally, in speech perception and recognition, the syllable has long been regarded as the perceptually most salient and robust unit of the acoustic stream [2]. From such a point of view, it seems natural to consider a hierarchical approach to speech recognition that first detects syllabic nuclei and proceeds with a coarse to fine analysis of the signal around these points for further segmentation, recognition, and learning. Indeed, approaches to speech recognition that are motivated by ideas in perception, phonetics, and phonology, have made attempts in this direction, like feature-based landmark detection [3], and our work is a contribution to this tradition. It is hoped that further work along this direction will lead to a phonetically motivated speech recognition system that will provide a viable alternative to the current tradition using HMMs and generic front-ends.

Prior Work in This Tradition: In fact, before statistical models dominated speech recognition research, the acoustics of syllables was widely analyzed and used into automatic detection. Weinstein used predominance of low frequency energy between 100Hz and 900Hz [4]. Kasuya and Wakita even employed a complicated combination of energy, back-to-total cavity volume ratio, front-to-back cavity volume ration and high-to-low frequency energy ratio to segment speech into vowel and non-vowel units [5]. These methods use features extracted from sound spectrogram, requiring to transfer the speech waveform in a function of amplitude and time into the signal spectrum represented in frequency and time.

The two important points of comparison for this paper are (i) Mermelstein's algorithm [6] which used convex hull algorithm on energy between 500Hz and 4kHz to segment speech into syllabic units; (ii) Howitt's work [7], incorporating ANNs into an energy-based acoustic vowel detector. Clearly, only energy is not sufficient enough to make reliable detection.

Our Central Result: Our central contribution is an algorithm for detecting syllabic nuclei from continuous speech. After carefully analyzing the acoustics of syllabic nuclei and other speech phonemes, our approach to syllable detection tries to use two reliable acoustic cues, periodicity and relevant energy. Using a modified version of a convex hull algorithm first introduced in this context by Mermelstein, we use a two step procedure to locate syllabic nuclei. The algorithm requires very simple computation. We do a detailed study of the performance of this algorithm with particular attention to how performance degrades with changes in speaker and channel characteristics including noise. We find that our algorithm is competitive with state-of-the-art HMM based approaches for this task in clean speech, is far more robust, requires less training data and computational resources, and is phonetically interpretable. In conjunction with other work in feature detection, we believe this may eventually lead to an overall speech recognition system. Noting that the vowels which form syllabic nuclei have formant transitions from beginning of the segmental duration to its end, we observe that the estimated vowel landmark is a point of relative stability in the vowel where its formant values may be closer to the target formant values for that vowel.

2. Syllable Detection Method

2.1. Periodicity

Due to continuous changes of vocal cords and tract, the speech signal, even for sound of vowels and sonorants, is actually quasi-periodic. Moreover, some vowels, especially short ones, are also affected by context phonemes, which makes them display aperiodic property. But generally, we can list the phoneme groups in order of periodicity from the most periodic to the least as vowels/sonorants, liquids/glides, nasals, whispers, fricatives/affricates,

and stops. This characteristic gives us a clue to separate syllabic nuclei from consonants. To evaluate periodicity, we use a modified autocorrelation function of time series.

Let $x_1, ..., x_n$ be observations of a time series, X_t. We can estimate the autocorrelation function of X_t using sample autocorrelation function, if X_t is a stationary time series, which is usually the case for speech signal. The sample autocovariance function is

$$\hat{\gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-|h|} (x_{t+|h|} - \bar{x})(x_t - \bar{x}), -n < h < n, \quad (1)$$

where $\bar{x}$ is the sample mean of $x_1, ..., x_n$.

Since the expected mean of speech signal is usually 0. Equation 1 can be rewritten as

$$\hat{\gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-h} x_{t+h} x_t, 0 \leq h \leq n-1, \quad (2)$$

where $\hat{\gamma}(0)$ is actually the total energy of the speech signal.

In order to compare values of phonemes with different pitch periods across an utterance, we normalized the autocorrelation function over the number of samples used for computation, as

$$P(h) = \frac{\hat{\gamma}(h)/(n-h)}{\hat{\gamma}(0)/n}, 0 \leq h \leq n-1. \quad (3)$$

The periodicity of a speech frame is then defined as the largest value among the peaks of the normalized autocorrelation function. Because the samples of sound signal are not independent, especially for adjacent samples, there will be fake peaks in the first few calculated values. And since the number of samples used to normalize autocorrelation to $P(h)$ is very small for large h, the resultant $P(h)$ will appear randomly at the end. In practice, only the peaks in the middle stable portion are considered for the periodicity value of a frame. In our implementations, only sample shifts from 40 to 240 are considered. It is actually pitch period from 2.5ms to 15ms.

The left top plot of Figure 1 shows the waveform of phoneme /eh/ in a frame of 25ms spoken by a female speaker from TIMIT data, and the left bottom shows the normalized autocorrelation. The periodicity can easily be found with value around 1.0. The right two are for phoneme /z/ spoken by the same speaker. The periodicity is around only 0.25, after cutting off the ends.

Analysis of some utterances in TIMIT training data shows that about 99% of syllabic nuclei have periodicity more than 0.5, and about 80% are even over 0.9, with average around 0.92, while other non-syllable consonants, like stops and fricatives, average at 0.45.

2.2. Relevant Energy

Energy of a speech frame can be represented as $log\hat{\gamma}(0)$, where $\hat{\gamma}(0)$ is calculated as in Equation 2. Because the amplitudes of sound can vary substantially from one utterance to another, the energy of each frame is normalized against the strongest frame energy in the utterance. In this way, the strongest frame in an utterance will have an relevant energy value of 0dB and others will have values below it.

For each phoneme, we select the frame with the largest energy to represent it and find that syllabic nuclei, vowels and sonorants, are averagely 15dB above other consonant groups. It can also be shown that more than 99% of syllabic centers have relevant energy over -50dB.

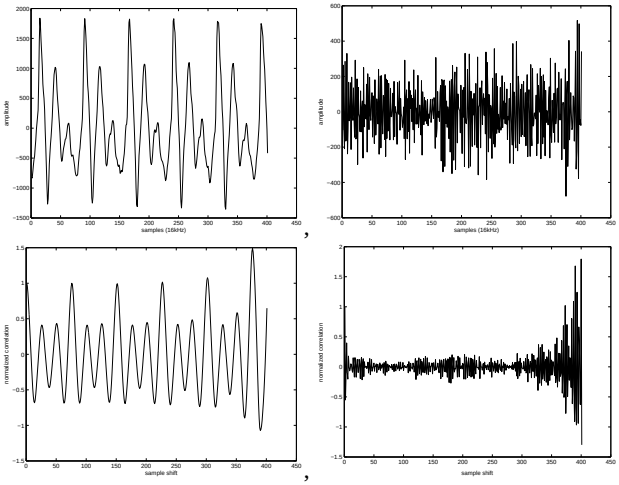

Figure 1: *Waveform and periodicity of /eh/ and /z/*

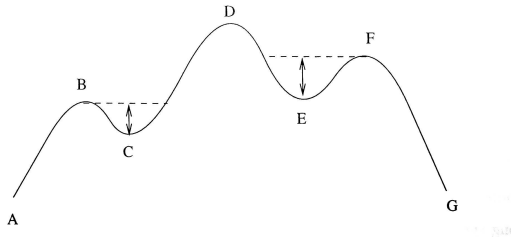

Figure 2: *Example of convex hull algorithm*

2.3. Landmark Detection

After collecting periodicity and relevant energy for each frame in an utterance, the landmarks of the syllabic nuclei are detected by two steps, described below.

Periodicity Segmentation A modified convex hull algorithm is used to segment an utterance into many segments. The basic algorithm first finds the maximal point of a series of data, and then constructs a convex hull, which is monotonically nondecreasing from the start of the segment to the peak point, and is monotonically non-increasing thereafter. Thus, at any time point, the value of the convex hull is at least as large as the value of the underline data. The differences at all time points between the convex hull and the data are calculated and the maximum of them is compared with a threshold value. If it is deeper than the threshold, this point serves as a boundary and the segment is divided into two subsegments. The construction and division process is recursively carried out onto the newly generated subsegments, until no differences are larger than the threshold value. For example, as in Figure 2, the convex hull is constructed as the dashed line with maximal point at D. The largest distance at E is compared with a threshold. If it is larger, the segment is divided into subsegment A-E and E-G. For segment E-G, the re-constructed convex hull will be the same as the underline segment, and then it cannot be further divided. As for segment A-E, the largest dip at C will be compared to the threshold to determine if C is a valid boundary.

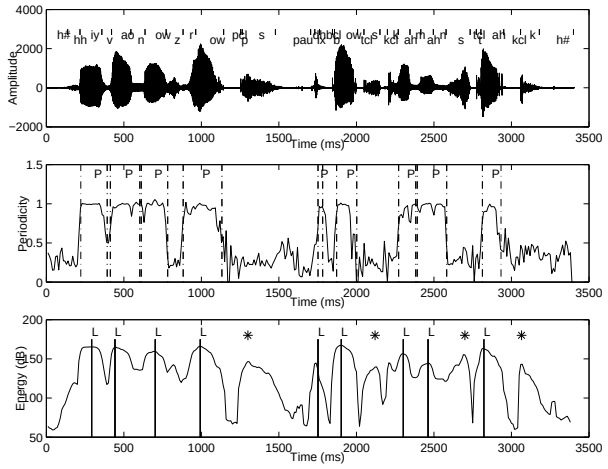

Figure 3: *Syllable detection example. The sentence is si2016, "Heave on those ropes; the boat's come unstuck."*

We apply the basic algorithm on periodicity to obtain all the segments. The values at the boundaries of each segment are further tested against a periodicity threshold and then each segment shrinks to center to form a periodic region and leave the two ends as aperiodic regions. An utterance is, thus, separated into periodic and aperiodic segments.

In order to remove stop closures and glottal stops, which are also highly periodic, we add another energy threshold test on each frame in the periodic segments.

Landmark Picking After obtaining periodicity segmentation, the basic convex hull algorithm is used on relevant energy to separate each periodic segment into several smaller sonorant regions. The energy peak of each region is located and marked as a syllabic center.

An example of the whole process is shown in Figure 3.The top graph is the amplitude plot against time from its waveform, with the phoneme script on the above. The middle is the periodicity with the resultant segmentation by the modified convex-hull algorithm. The segments with label "P" are the final periodic segments after shrinking from the convex hull segments. And the bottom figure is the syllable landmark output from the detector. The landmarks are located at the energy peaks of each periodic segments. Without periodicity segmentation, there would be false landmarks indicated by asterisk, which might be output by the unmodified convex hull algorithm.

3. Performance Testing

Based on analysis of several utterances in TIMIT training dataset, we use a typical set of parameters in our baseline syllable landmark detector, shown in Table 1.

Because TIMIT does not provide syllable information, we consider both vowels and sonorants (/el/, /em/, /en/, /eng) as syllabic nuclei. There are a total of 1344 utterances and 17190 syllabic nuclei in TIMIT test dataset, excluding SA1 and SA2 utterances. Howitt's detector is tested only on the vowels of 375 utterances in TIMIT test dataset. The performance, compared with other models, is shown in Table 2. The HMM-based Sphinx systems are developed by CMU. The referred Sphinx2 model is a semi-continuous model based on 5-state Bakis HMM topology, using 6000 context-dependent tied states for all the triphones of the 40 base phones. The Sphinx3 model is a continuous model with 8 Gaussians per state, based on 3-state HMMs with no skips and containing 6000 senones. No language models are involved.

Table 1: *Parameter setting for the baseline detector*

Parameter	Value
frame size	400 samples (25ms)
frame shift	160 samples (10ms)
energy threshold	50dB below the maximum
periodicity peak-to-dip threshold	0.7
energy peak-to-dip threshold	4.5dB

Table 2: *Performance of detection*

	Accuracy	Del. Error	Ins. Error	Total Error
Baseline	81.6	18.4	10.9	29.3
Sphinx 2	84.5	15.5	22.3	37.8
Sphinx 3	89.1	10.9	25.7	36.6
Howitt	75.5	24.5	13.8	38.3

A landmark detection of the baseline detector is counted correct, if it is located within a syllabic segment, provided by TIMIT phoneme boundary information. Because the outputs of Sphinx systems are phoneme segments, we transfer them into landmarks by putting a landmark at the middle of each syllabic segment, and then apply the same testing method. Both the deletion and insertion error rates are calculated against the number of syllables in the testing data. The total error rate is the sum of the deletion and insertion error rates. Clearly, the overall performance of our baseline detector is comparable to the complicated Sphinx systems, with even less total error rate. It also outperforms Howitt's detector using ANNs in both accuracy and error rates.

4. Robustness

In order to test how robust our detector and other systems are, two datasets are used, NTIMIT and TIMIT with additive noise.

4.1. NTIMIT and White Noise

There are two major differences between TIMIT and NTIMIT data, (i) more noise in NTIMIT speech with 25dB SNR while TIMIT has about 40dB SNR, (ii) greatly reduced spectral energy above 3.5 kHz in NTIMIT, due to telephone channel limitation. Among the released NTIMIT data, 8 speech files are incomplete, so the test data contains 1339 utterances with 17117 syllables.

In order to test degradation to additive noises, two kinds of noise are added to the TIMIT speech wave forms, global white noise and local white noise.

Global noise is added by sampling from a zero mean random distribution. The variance of the distribution is set depending upon the level of global signal-to-noise ratio (SNR) we wish to obtain. We add 30dB, 20dB, 10dB and even 0 dB SNR global noise to the TIMIT speech.

Local white noise is added in way of Schroeder noise. The

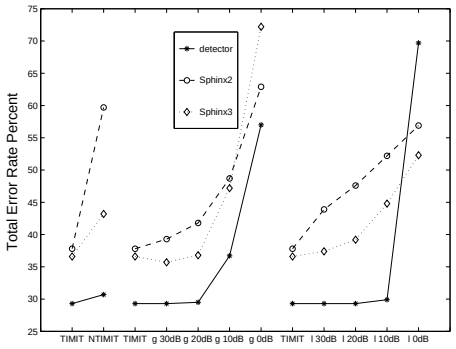

Figure 4: *Degradation of total error rates*

clean speech signal $x(n)$ is corrupted into

$$y(n) = x(n)[1 + \epsilon\eta(n)], \tag{4}$$

where $\eta(n)$ takes on values +1 and -1 with equal probability at each n, and any two $\eta(n)$ are independent. Therefore, the noise will have flat power spectrum. Furthermore, the signal-to-noise ratio at each time n (local SNR) is given by $20log(1/\epsilon)$. In this model, the total additive local noise $\zeta(n) = \epsilon x(n)\eta(n)$ scales as a function of signal energy at each point in time.

4.2. Degradation Test

We test our baseline detector and the two Sphinx models on the NTIMIT and noise-added TIMIT test datasets. Our detector uses the same parameter setting as Table 1. The Sphinx models are trained on TIMIT training data and not re-trained or adapted to NTIMIT and noise. Figure 4 is the degradation of total error rates. From the result, our detector is more stable and outperforms the HMM based systems in all environments, except for 0dB SNR local noise. The Sphinx systems even degrades right after small noise is added or the situation changes, like in 30dB noise and NTIMIT environments.

5. Application to Vowel Classification

Speech signal is a continuous sequence of individual sounds and is often in a transitory state, the dynamics of which seems to be well represented by HMMs. However, because the direction of the movement of the articulators usually approximates the target configuration of phonemes, we believe that there are still some time points describing the phonemes approximately well, especially for some long phonemes, like vowels. Then the signals around these time points can be used as reliable sources to classify phonemes. The landmarks detected in the previous sections can be such kind of reliable points.

Actually, we analyze vowel positions statistically and find that the syllable landmarks can be used as good approximation to the middle frames of vowels, which have much less variation than the starting and ending frames. Shown in Table 3 are speech variations at energy peak points of vowels and across the whole vowels. The statistical data are the second formants of vowel /ae/ spoken by eight male speakers, and vowel /iy/ spoken by eight female speakers. We construct statistical hypothesis tests and reject in either case the assumption that the variations are equal at peaks and across the phonemes.

Table 3: *Variations of second formants of /ae/ and /iy/*

	#(all)	mean	var	#(peak)	mean	var
/ae/	447	1648.49	215.31	61	1612.28	157.36
/iy/	649	2348.32	333.30	140	2299.73	294.75

Using the data at the landmarks extracted from the detector described in previous sections, we can build our vowel classifier based on Support Vector Machines. Considering the dynamics of speech signals, we use more than one frames of the vowel features at the landmarks to train our classifiers. Different groups of vowel features are tried and we obtain similar results. One of the feature groups discussed in this section consists of the energy ratios of 15 mel scale bands under 4000Hz over the total from the adjacent 3 frames. And the other one is the same cepstral feature used by the Sphinx systems. The classification results are shown in Table 4, without diphthongs. The comparable performance implies that the syllable landmarks are indeed reliable for classification.

Table 4: *Performance of vowel classification*

	mel	cepstral	Sphinx 2	Sphinx 3
Accuracy	48.0%	48.5%	46.1%	50.9%

6. Conclusions

We have developed a syllable detector using basic features of periodicity and relevant energy. These two features are relatively robust across speakers and utterances. Their stability and reliability are demonstrated in the robustness tests and vowel classification. Clearly, thorough analysis of signal characteristics and careful selection of distinctive features can overcome some problems coming from statistical models and improve speech recognition.

7. References

[1] Saffran, J. R., Senghas, A., and Trueswell, J. C. (2001). The acquisition of language by children. Proceedings of the National Academy of Sciences, 98, 12874-12875.

[2] Greenberg, S. (1999). Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation. Speech Communication v. 29 (1999), 159-176

[3] Stevens, Kenneth N., Manuel, Sharon Y., Shattuck-Hufnagel, Stefanie, Liu, Sharlene (1992). Implementation of a model for lexical access based on features. ICSLP-1992, 499-502.

[4] Clifford J. Weinstein, Stephanie S. McCandless, Lee F. Mondshein, and Victor W. Zue. A system for acoustic-phonetic analysis of continuous speech. IEEE Trans. ASSP, 23(1), 1975.

[5] Hideki Kasuya and Hisashi Wakita. An approach to segmenting speech into vowel- and nonvowel-like intervals. IEEE Trans. ASSP, 27(4), 1979.

[6] Paul Mermelstein. Automatic segmentation of speech into syllabic units. JASA, 58(4), 1975.

[7] Andrew Wilson Howitt. Automatic syllable detection for vowel landmarks. PhD thesis, MIT, 2000.