INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Weakly Supervised Syllable Segmentation by Vowel-Consonant Peak Classification

Ravi Shankar[1], *Archana Venkataraman*[1]

[1]Department of Electrical and Computer Engineering, Johns Hopkins University

rshanka3@jhu.edu, archana.venkataraman@jhu.edu

Abstract

We present a novel approach for blind syllable segmentation that combines model-based feature selection with data-driven classification. In particular, we learn a function that maps short-term energy peaks of a speech utterance onto either the vowel or consonant class. The features used for classification capture spectral and energy signatures which are characteristic of the phonetic properties of the English language. The identified vowel peaks subsequently act as the nucleus of our syllable segments. We demonstrate the effectiveness of our proposed method using nested cross validation on 400 unique test utterances taken randomly from the TIMIT dataset containing over 5000 syllables in total. Our hybrid approach achieves lower insertion rate than the state-of-the-art segmentation methods and a lower deletion rate than all the baseline comparisons.

Index Terms: Syllable segmentation, Gaussian process regression, peak detection, random forest classification

1. Introduction

Syllables play a crucial role in speech articulation and perception [1] by providing insight into the rhythmic aspects of speech. It was shown by [2] that syllables are a natural unit for measuring the quality of interaction in dialogue systems. Syllables also have the potential to be useful for emotion recognition and speech rate estimation as shown in [3] and [4, 5], respectively. Despite these advantages, syllables have received very little attention in the speech community. The main challenge in developing syllable based models is the huge amount of variability of these acoustic units across any given language. For example, Hidden Markov Models (HMM) are easy to train at phoneme level because of their limited number (44 - 46 unique phonemes in English), and the large amount of training data available for each phoneme. In contrast, there are over $10,000$ syllables in the English language, so collecting and annotating sufficient training data for each of these variants is an extremely daunting task [6]. Furthermore, natural differences in articulation also leads to copious variations in the duration and the energy profile of a syllable. The main focus of this paper is to develop an automated method to segment a speech utterance into its constituent syllables with low model complexity.

1.1. Relation to prior work

The problem of blind syllable segmentation is closely tied to that of localizing landmarks in speech (vowels, semi-vowels and fricatives) as a proxy for speaking rate [7]. The works of [8, 9] carry these ideas one step further by using such landmarks to guide blind syllable segmentation. Specifically, the Praat script proposed in [8] uses the highest peak in the loudness contour as the syllable nuclei. In contrast, the Syll-o-Matic approach of [9] fuses the loudness and voicing measure to generate a distribution over the boundary and landmark locations.

An alternative approach for syllable segmentation was proposed by Mermelstein [10] as an iterative procedure. It computes the short-term energy and then sequentially draws syllable boundaries based on relative local maximas in the difference between the energy signal and its convex hull. This procedure is repeated for each smaller segment until some stopping criterion is reached. The work of [11] modified the pre-processing step of the Mermelstein's algorithm to use the energy present in just the fundamental frequency and the first formant for segmentation. Another method proposed by Villing et al. [12] uses a 70 db equal loudness filter to selectively enhance specific frequency ranges based on models of human auditory perception. The resulting signal is further decomposed into three channels. The syllable segmentation is performed based on the onset velocity extracted from the envelope of these channels. Recent works on syllable segmentation involves using deep neural networks to identify nucleus based on their sonority profile [13].

Our proposed method extends the prior work by learning a data-driven model for the syllable nuclei identification in a supervised fashion. Our strategy is to classify local maximas (i.e. peaks) in the short-term energy of the speech utterance as belonging to either a vowel or consonant phoneme. This classification is based on features extracted from the energy signal and from the spectrum of original speech signal. We use an ensemble classifier to ensure robustness. The identified vowel peaks act as the central element of the syllable segments. The segmentation boundaries are then drawn based on a learned minimum temporal separation threshold. We demonstrate the results of our method on the standard TIMIT [14] dataset. Our proposed algorithm achieves lower overall term error rate than the baseline algorithms. This validation shows that the hybrid approach, though simple, is effective for syllable segmentation.

2. Segmentation via Energy Peaks

Fig. 1 outlines our method. Given a speech utterance, we extract its energy signal and smooth it by applying the Gaussian process regression. We then classify points of local maximas in the smoothed energy contour as belonging to either a vowel or a consonant phoneme. Peaks that are identified as vowels will act as the nucleus of our segmentation algorithm.

2.1. Extraction of Short-Term Energy Signal

Our pre-processing steps to extract the energy signal are similar to what has been described in [11]. The original speech utterance $X(t)$ is first passed through a second-order Butterworth lowpass filter with a cut-off frequency of 1200 Hz. The high cut-off frequency ensures that the filtered signal retains all of its energy from the first formant along with some component of the second formant information. The filtered signal is then sample-wise squared to obtain the instantaneous power.

The derived power signal is lowpass filtered again with a second-order Butterworth filter having a cut-off frequency of

http://dx.doi.org/10.21437/Interspeech.2019-1450

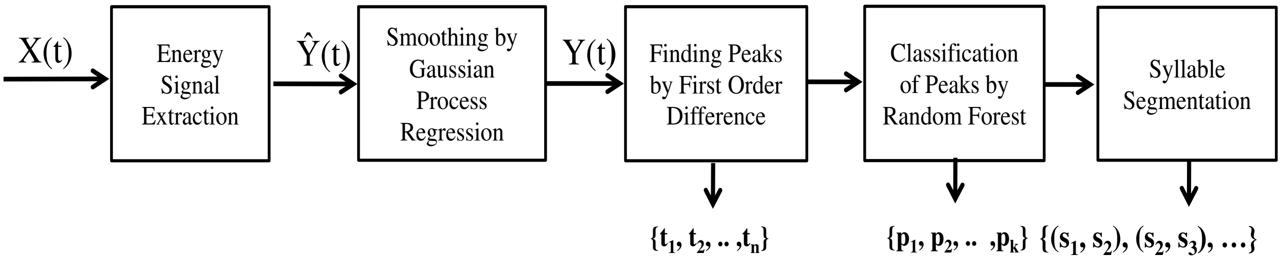
X(t)
Energy Signal Extraction
Ŷ(t)
Smoothing by Gaussian Process Regression
Y(t)
Finding Peaks by First Order Difference
Classification of Peaks by Random Forest
Syllable Segmentation
{t1, t2, .. ,tn}
{p1, p2, .. ,pk}
{(s1, s2), (s2, s3), ...}

Figure 1: *Flow-chart of proposed segmentation algorithm. $X(t)$ is the input speech utterance, $\hat{Y}(t)$ is the energy contour and $Y(t)$ denotes the smoothed energy contour. Time stamps $\{t_1, t_2, .., t_n\}$ represents the local maximas in the signal $Y(t)$ while $\{p_1, p_2, .., p_k\}$ are just the vowel peaks obtained from classification. Variables $\{(s_1, s_2), (s_2, s_3), ..\}$ are the derived syllable segments.*

15 Hz. This choice accounts for the perception model for speech, which is dominated by the smoothly varying envelope of the signal [15]. The final curve $\hat{Y}(t)$ is denoted as the short-term energy signal (i.e, loudness function) described in [10].

2.2. Smoothing and Peak Detection

The loudness function $\hat{Y}(t)$ exhibits noisy ripples, which severely confound our peak detection. This problem is addressed by estimating a smooth envelope of the energy signal using Gaussian process regression [16]. Given a set of N sampled time points $T_N = \{t_1, t_2, .., t_N\}$ and target function values $Y_N = \{y_1, y_2, .., y_N\}$, we can compute the posterior distribution of the smooth envelope $f(\cdot)$ at new time points t' according to the rule of conditional probability:

$$P(f(t')|T_N, Y_N, t') = \frac{P(f(t'), Y_N|T_N, t')}{P(Y_N|T_N)} \tag{1}$$

The term in the denominator is the partitition function:

$$P(Y_N|T_N) = \int P(Y_N|T_N, f(\cdot))\; P(f(\cdot))\; df \tag{2}$$

where $f(t)$ is a random process with mean zero and covariance function $K(t, t')$ i.e. $f(t) \sim N(0, K(t, t'))$. We use a radial basis function (RBF) to model local similarities in the speech:

$$K(t, t') = \alpha \exp \frac{-||t - t'||^2}{2\tau^2} \tag{3}$$

Here, the parameters α and τ control the kernel amplitude and smoothing scale, respectively. They are learned by minimizing the negative log likelihood of Y_N with respect to α and τ:

$$\begin{aligned} \log\left(P(Y_N|T_N)\right) &= \frac{-1}{2} \log|\mathbf{K} + \sigma^2 I| \\ &- \frac{1}{2} Y_N^T (\mathbf{K} + \sigma^2 I)^{-1} Y_N - \frac{N}{2}\log(2\pi) \end{aligned} \tag{4}$$

where, $\mathbf{K}$ is now an $N \times N$ Gram matrix computed between the finite observations. Fig. 2 shows the result obtained by applying Gaussian process regression on a noisy energy signal.

Notice that there is a trade-off between smoothing the energy curve and preserving the vowel peaks. As the degree of smoothness is increased, the number of spurious peaks decrease, but we are more likely to lose some less prominent intensity peaks. Empirically, we find that uniformly sampling the energy signal at anywhere between 60-66 Hz, which corresponds to roughly 200 points per 2-3 sec utterance, leads to a good balance between vowel miss rate and smoothness.

Finally, we use a first order difference to identify local maximas, i.e., peaks in the smoothed energy contour $Y(t)$.

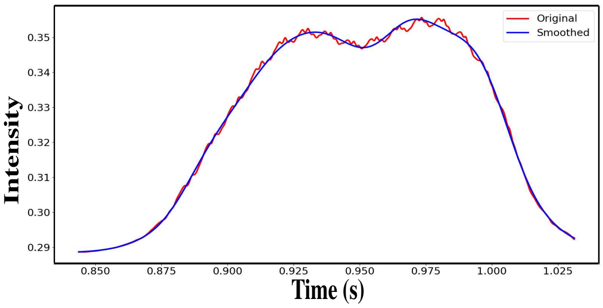

Figure 2: *Illustration of Gaussian Process smoothing. The blue curve is the smooth approximation of the original (red) noisy signal. It also retains the trend present in original signal.*

2.3. Features Extraction

We derive four classes of hypothesis-driven features for the vowel/consonant classification. These features are:

Peak Intensity: Since voiced sounds require a significant vibration of the vocal cords, we expect vowel phonemes to have higher energy than most consonants. Hence, we select the intensity of each local maxima in the short-term energy signal.

Slope of Local Maxima: Vowel peaks typically have a more gradual onset and decay than harsh consonants such as /p/, /k/ and /t/. We calculate the time for the signal to decay from the peak value to half this maximum as a proxy for its slope. It helps to discriminate between vowels and plosives.

Number of Local Maximas in a Small Window: Diphthongs such as /ou/ and /ow/ generate multiple energy peaks that are closely spaced to each other with approximately the same intensity. Hence, we use number of local maximas in a small window of 160 ms centered at a given peak as our third feature.

Energy Features: In addition to the peak features, spectral features in the original signal are particularly important for disambiguating nasals (/m/, /n/) and liquid consonants (/l/, /r/), which have a very similar intensity profile to vowels. We extract the following spectral attributes to improve detection accuracy:

- Energy in the $100 - 200$ Hz band
- Energy in the $200 - 400$ Hz band
- Energy in the $400 - 600$ Hz band
- Energy in the $0 - 600$ Hz band

Finally, the energy in the 20 Mel frequency bands is also used to improve the consonant detection performance. In total, we have 27 features to train our vowel/consonant peak classifier.

2.4. Peak Classification

The peak labels for each training utterance are generated via the Penn forced alignment package [17]. It takes a speech file and its transcription as an input and returns the time intervals of each phoneme in the utterance. Peaks in the energy signal that are located in the vowel phoneme intervals are then labeled as vowel peaks and likewise for consonant peaks.

Random forest is an ensemble classifier that fits multiple decision trees on the same data by bootstrapping [18]. Each decision tree in the ensemble is trained on a random subset of utterances from the training set. During testing, each decision tree independently predicts the class for the test sample, and a majority voting scheme is used to assign the final label. The decision tree itself is like a flow-chart, where each non-leaf node denotes a query on a single feature, the corresponding branch represents the outcome of that query, and each leaf node specifies a class label. We used Gini Impurity [19] as the feature selection criterion at each node. We also restrict the maximum depth of each tree to seven and the number of trees to 50 for good generalization. We did not find any performance gain by moving from a random forest to a neural network architecture.

2.5. Syllable Segmentation

The inputs $y(t)$, $\{p_1, p_2, ..., p_k\}$ and θ correspond to the smoothed energy contour, the temporal location of vowel peaks obtained from the random forest classifier, and the minimum separation threshold, respectively. At each vowel peak $p \ \epsilon \ \{p_1, p_2, ..., p_k\}$, we make a decision whether or not to draw a syllable boundary based on its distance from the last boundary location. If the difference is greater than the threshold ($\theta = 150 \ ms$, in our study) then we identify the lowest minima near the midpoint of next vowel peak and the current one. This point becomes our next syllable boundary. Further, we constrain that the end point of the speech utterance is the last syllable boundary segment regardless of separation threshold.

3. Experiments and Results

3.1. Evaluation Dataset

We evaluate our algorithm on the 2300 utterances from TIMIT corpus. We use 1700 utterances to train our random forest classifier. The forced alignment is used to generate vowel/consonant labels for the training data. The remaining 500 utterances are syllabified using rule based syllabifier, $tsylb$ [20] to create ground truth segments for testing. Tsylb is a standard package used for ground truth evaluation in recent syllable segmentation work [13]. Of these 500 utterances, we use 100 as our validation set to learn the model parameters for both our method and each of the baselines. The final 400 utterances are used for the primary evaluation across all algorithms.

3.2. Baseline Methods

We compare our peak based segmentation algorithm with two different types of blind segmentation procedures. The first class performs a direct syllable segmentation based on the envelope of a filtered version of the speech utterance. Here, we have implemented the Mermelstein [10] and Villing et al. [12] algorithms, which are the current state-of-the-art. The second class identifies landmarks in the energy profile as the basis for segmentation [8, 9]. In particular, we have implemented the Syll-o-Matic [9] and the Praat script [8] methods described in prior work. Since the Praat script method only returns the syllable nuclei, we implement our own boundary procedure to create the syllable segments for a direct comparison.

Table 1: *Performance of syllable segmentation. Comparison with envelope based (top) and landmark based (bottom).*

Algorithm	Insert	Delete	TER	Overlap
		Envelope Based		
Mermelstein [10]	0.16	0.20	0.36	0.67
Villing [12]	0.14	0.22	0.36	0.66
Proposed	**0.11**	**0.19**	**0.30**	**0.69**
		Landmark based		
Praat Script [8]	**0.03**	0.39	0.42	0.52
Syll-o-matic [9]	0.07	0.30	0.37	0.60
Proposed	0.11	**0.19**	**0.30**	**0.69**

Table 2: *Comparison of vowel peak detection performance.*

Algorithm	True Detection	False Alarm	AUC
Praat Script	0.522	**0.114**	0.71
Syll-o-matic	0.496	–	–
Proposed	**0.94**	0.30	**0.82**

3.3. Blind Syllable Segmentation Performance

The segments obtained by each algorithm are compared against their corresponding ground truth transcription. An overlap of greater than 50% with the ground truth is considered as a valid syllable segment. Fraction of insertion, deletion, and overlap are calculated according to [12] and are averaged over the entire test dataset. Table 1 summarizes and compares the performance of baseline methods with our proposed algorithm.

Our proposed method has lower error rates and higher overlap than both envelope based methods. This gain in the performance can be attributed to our carefully selected features that differentiate between vowels and high-intensity consonants. In contrast, the envelope based methods consider just the intensity profile and are frequently confounded by nasal (/n/) and liquid consonants (/l/, /r/). Another advantage of our method is the minimum separation threshold between the vowel nuclei. This constraint further eliminates spurious segments.

In comparison to landmark based approaches, our proposed method has substantially lower deletion and term error rates. In fact, as shown in Table 2, the landmark based approaches detect on average, only half of the vowel peaks in a given utterance. Said another way, landmark detection using just the loudness and sonority functions is no better than a coin flip. This lackluster performance is due to the overlapping distributions between the intensity of vowel and consonant phonemes. Once again, our careful feature specification helps us to disambiguate these classes. Moreover, the random forest can extract complex decision boundaries in the underlying feature space. When combined, our vowel peak detection rate is well over 90% with the Area Under Curve (AUC) score of over 80%. Hence, a combination of hypothesis driven feature selection and machine learning classification has the potential to outperform heuristic landmark detection. One thing to note is that the Syll-o-Matic landmarks confound the vowels, semi-vowels and fricatives, so we cannot compute the false alarm rate or AUC for this method.

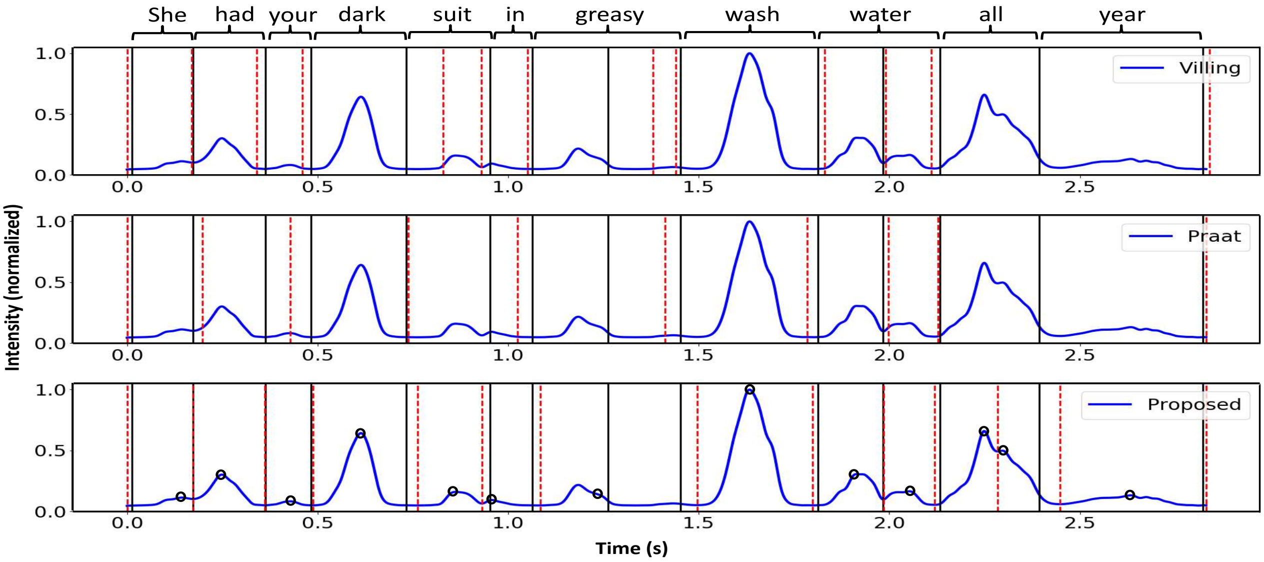

Figure 3: *Illustration of syllable boundaries obtained by Villing (top), Praat Script (middle) and proposed peak detection (bottom). The solid black lines denote the ground truth boundaries and the dotted lines are the segments obtained by the corresponding algorithm. Vowel peaks identified by random forest are marked by black circles in the bottom figure.*

Fig. 4 highlights the robustness of our peak detection approach. In particular, note that our method and the Mermelstein algorithm have noticeably higher median accuracy that the other three baselines. Furthermore, when compared to Mermelstein, our median accuracy is slightly better with a tighter interquartile range. Consequently, our hybrid learning strategy achieves consistent performance across a diverse set of testing utterances. A one sample t-test for each pair-wise difference (shown in Fig 4) results in p-values less than 0.05 which validates our claim.

Fig. 3 shows an example segmentation obtained by an envelope method (Villing), a landmark based approach (Praat script) and our proposed algorithm. Notice that, the Villing method creates many short syllable segments because it is sensitive to the onset characteristics in its estimated envelopes, and it does not impose a temporal separation constraint. For instance, the Villing method places a boundary right in the middle of the word $suit$, where the onset happens to be in the loudness contour. In contrast, the Praat script has the opposite problem; it misses a lot of the true boundary locations including the one between the words $suit$ and in. This performance is to be expected, given its poor syllable nuclei detection rate. Our approach recovers the boundary between these two words, since the vowel peaks have been classified correctly by the random forest; this example highlights the strength of our model-driven feature set. A final point to observe here is that all three of the methods miss the boundary located to the left of 1.5 second mark. This particular boundary is hard to detect because it does not lie at any local minima in the intensity contour. Moreover, the peak corresponding to the nuclei of the second syllable $s{-}iy$ in the word $greasy$ (g-r-**iy**-s-**iy**) is very ambiguous. It stays flat for its entire duration. One possible explanation for this behavior is that the $s - iy$ was not highly emphasized during its utterance. This scenario illustrates why syllable segmentation is a difficult problem in speech processing. Nonetheless, our proposed algorithm greatly outperforms both the state-of-the-art envelope methods and other landmark-based approaches.

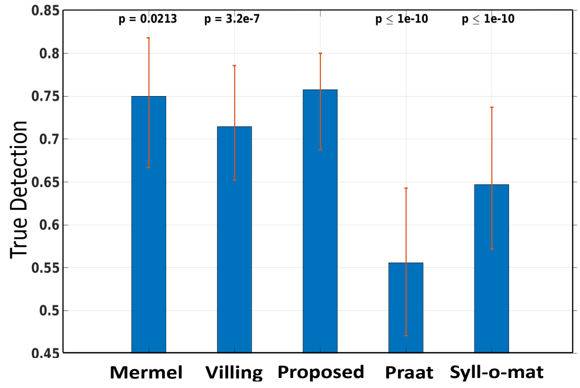

Figure 4: *Bar plot of true detection across test utterances. The number on top of each bar plot shows the p-value obtained from two sample t-test with proposed algorithm.*

4. Conclusion

We have developed a method to automatically segment syllables from continuous speech based on identifying vowel peaks in the short-term energy contour of a signal. The proposed algorithm iterates over vowel peaks identified from supervised learning task and selectively places a syllable boundary between them if separated by more than 150ms. This simple approach outperformed the existing algorithms for syllable segmentation on the dataset used for evaluation. Specifically, the segmentation performance improvement came from the careful feature selection that was made for disambiguation of vowel peaks from the consonants in the energy contour. It lead to significant improvements in all the performance metrics considered for syllable segmentation over the existing state-of-the-arts. In future, our algorithm can be implemented in a streaming fashion by processing intervals of the speech signal in batches.

5. References

[1] S. Greenberg, "Speaking in shorthand - a syllable-centric perspective for understanding pronunciation variation," *Speech Communication*, vol. 29, no. 2, pp. 159–176, 1999.

[2] N. Fujiwara, T. Itoh, and K. Araki, “Analysis of changes in dialogue rhythm due to dialogue acts in task-oriented dialogues,” *International Conference on Text, Speech and Dialogue*, vol. 4629, pp. 564–573, 09 2007.

[3] A. Origlia, F. Cutugno, and V. Galatí, “Continuous emotion recognition with phonetic syllables,” *Speech Communication*, vol. 57, pp. 155–169, Feb. 2014.

[4] 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, Nov 2007.

[5] C. Yarra, O. D. Deshmukh, and P. K. Ghosh, “A mode-shape classification technique for robust speech rate estimation and syllable nuclei detection,” *Speech Communication*, vol. 78, pp. 62 – 71, 2016.

[6] J. B. Pierrehumbert, “Syllable structure and word structure: a study of triconsonantal clusters in English BT - Phonological structure and phonetic form: Papers in Laboratory Phonology III,” *Phonological structure and phonetic form: Papers in Laboratory Phonology III*, pp. 168–190, 1994.

[7] Y. Zhang and J. R. . Glass, “Speech rhythm guided syllable nuclei detection,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2009.

[8] N. H. de Jong and T. Wempe, “Praat script to detect syllable nuclei and measure speech rate automatically,” *Behavior Research Methods*, vol. 41, no. 2, pp. 385–390, 2009.

[9] N. Obin, F. Lamare, and A. Roebel, “Syll-o-matic: An adaptive time-frequency representation for the automatic segmentation of speech into syllables,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 6699–6703.

[10] P. Mermelstein, “Automatic segmentation of speech into syllabic units.” *The Journal of the Acoustical Society of America*, vol. 58, no. 4, pp. 880–883, 1975. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa

[11] A. W. Howitt, “Automatic Syllable Detection for Vowel Landmarks,” *Dissertation Abstracts International, B: Sciences and Engineering*, vol. 62, no. 6, 2001.

[12] R. C. Villing, J. M. Timoney, T. E. Ward, and J. K. Costello, “Automatic Blind Syllable Segmentation for Continuous Speech,” in *ISSC, Belfast*, 2004.

[13] C. Landsiedel, J. Edlund, F. Eyben, D. Neiberg, and B. Schuller, “Syllabification of conversational speech using bidirectional long-short-term memory neural networks,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 06 2011, pp. 5256 – 5259.

[14] J. S Garofolo, L. Lamel, W. M Fisher, J. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, “Timit acoustic-phonetic continuous speech corpus,” in *Linguistic Data Consortium*, 11 1992.

[15] J. M. F. Rob Drullman and R. Plomp, “Effect of temporal envelope smearing on speech perception,” *The Journal of the Acoustical Society of America*, 1994.

[16] D. J.C. MacKay, “Introduction to gaussian processes,” *NATO Adv Stud Inst Ser F Comput Syst Sci*, vol. 168, 01 1998.

[17] J. Yuan and M. Liberman, “Speaker identification on the SCOTUS corpus.” *Proceedings of Acoustics ’08*, 2008.

[18] P. F. et. al., “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[19] T. M. Mitchell, “Machine learning and data mining,” *Commun. ACM*, vol. 42, no. 11, pp. 30–36, 1999. [Online]. Available: http://doi.acm.org/10.1145/319382.319388

[20] W. Fisher, “Tsylb syllabification package.” *https://www.nist.gov/file/65961*, 1996.