Unsupervised word discovery from speech using automatic segmentation into syllable-like units

Okko Räsänen[1,2], *Gabriel Doyle*[2], *Michael C. Frank*[2]

[1] Aalto University, Department of Signal Processing and Acoustics, Finland
[2] Stanford University, Language & Cognition Lab, California

okko.rasanen@aalto.fi, gdoyle@stanford.edu, mcfrank@stanford.edu

Abstract

This paper presents a syllable-based approach to unsupervised pattern discovery from speech. By first segmenting speech into syllable-like units, the system is able to limit potential word onsets and offsets to a finite number of candidate locations. These syllable tokens are then described using a set of features and clustered into a finite number of syllable classes. Finally, recurring syllable sequences or individual classes are treated as word candidates. Feasibility of the approach is investigated on spontaneous American English and Tsonga language samples with promising results. We also present a new and simple, oscillator-based algorithm for efficient unsupervised syllabic segmentation.

Index Terms: zero-resource speech processing, unsupervised learning, segmentation, syllabic segmentation, time-domain analysis, speech perception, speech rhythm

1. Introduction

Learning to recognize words from speech is a challenge that human infants learn to solve during their first years of life. Related to this fundamental unsupervised learning challenge, so-called "zero-resource" speech processing has received recent interest in the speech engineering community as the access to labeled training data required for typical supervised machine learning is limited in many situations [1]. Although children's early word learning likely takes place in the context of rich social interactions rather than from pure speech input (e.g., [2]), learning from acoustic input alone provides insights into what information a human learner can and cannot obtain from the acoustic signal, as well as a starting point for speech recognition in underdocumented languages.

Addressing both scientific and technical aspects of zero-resource speech processing, the present paper investigates a syllable-based approach for unsupervised discovery of word-like patterns from continuous speech. In contrast to earlier work, which has used dynamic time-warping (DTW) [3-6], high-order Markov-approximations [7], HMMs [8], or spectrotemporal parametric models [9] for finding recurring fragments of audio, the present work attempts to bootstrap the learning process with the help via the rhythmic properties of speech.

The current system has three main stages: 1. discovery of a collection of syllable-like units (from now on just *syllables*) using unsupervised amplitude envelope-based segmentation, 2. representation of these segments in a feature space, and 3. treatment of these syllables and their recurring combinations as potential words. By starting with the syllable segmentation, the system is able to limit the potential onsets and offsets for patterns of interest, and also leading to computationally efficient temporal normalization between different acoustic tokens. Because of these choices, the present approach is efficient, enabling super-real-time processing of speech data with modest computational resources. We also present a novel and efficient oscillator-based method for automatic syllabification of speech that is inspired by neural models of speech perception.

1.1. On the role of syllables in speech perception

Syllables and syllabic stress patterns have been long considered to be central to language acquisition and speech perception (e.g., [10-12], see also [13]). Rhythmic structure is more prominent in infant-directed speech, perhaps facilitating early word segmentation (e.g., [14]). Syllables also coincide with the time-scale at which human hearing integrates acoustic input over time, a scale within which there are strong statistical temporal dependencies in the signal (e.g., [15-17]; see also [18] and references therein).

Syllabic structure appears important at a neural level, as well. Neural oscillations in auditory cortical areas are synchronized to the amplitude envelope of speech; this phase-locking may be required for speech comprehension [19]. Conceptual models of speech perception by Ghitza [20] and Giraud and Poeppel [21] explicitly posit that these so-called theta-range oscillations at the syllabic rate ($\approx$ 4–7 Hz) define "packages" of information within which more-detailed phonetic information is analyzed and integrated. Together with the common assumption that coarticulatory effects are smaller across than within syllables, the amount of empirical evidence and theoretical attractiveness make syllable a potential candidate for being a basic structural unit of speech.

Despite this body of evidence, the majority of existing computational work on human language processing and word learning has focused on the analysis at the phone (or phoneme) and word levels. Similarly, syllables are rarely utilized in the mainstream speech technology applications (but see, e.g., [22-24]). One reason for the absence of syllable-centered studies is likely because the entire concept of a "syllable" is elusive. For instance, syllabification of spoken or written English is notoriously difficult, and unanimous definition of a syllable in the articulatory domain is also far from trivial.

An additional issue for mainstream, supervised systems and even phonetic analysis of speech is that once accurate phonetic transcriptions of the input are available, analysis at the syllabic level does not necessarily provide much new information. The segmentation and sub-word representation problems have already been solved, if not in terms of phones, then at least using context-dependent phone models. However,

when a detailed linguistic coding of the input is not available, rhythmic structure of speech associated with syllables could be useful for bootstrapping the language learning process in both humans and zero-resource computational systems.

In the present work, we investigate the feasibility of this type of syllable-timed speech processing. In this context, we define syllables as segments of speech that are characterized by notable rhythmic increases and decreases in signal amplitude at the time-scale of approximately 2-10 Hz. This work is part of the Interspeech-2015 Zero-Resource Speech Challenge [25] where the goal is to discover linguistic units from conversational speech in a purely unsupervised manner.

2. Methods

An overview of the processing stages is shown in Fig. 1. First, the incoming speech is segmented into syllable-like units using the amplitude envelope of the signal. Each segment is then described with a fixed-length feature vector, effectively imposing time normalization on the segmental units. All individual segments are clustered into a number of discrete categories based on the similarity of the features across the segments. Finally, potential patterns of interest are extracted by searching for recurring segment-combinations or frequent individual segments (n-grams of different orders).

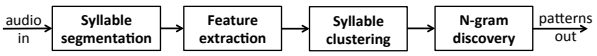

Figure 1: *A block schematic of the processing pipeline.*

2.1. Syllable segmentation algorithms

Three different algorithms for syllabic segmentation were investigated: 1) an algorithm proposed by Villing et al. [26] and used in the AuToBI toolbox [27] (from now on, *VSeg*), 2) a simple amplitude envelope minima detector, and 3) a novel neurophysiology-inspired damped oscillator model.

2.1.1. VSeg algorithm

The basic idea in VSeg is to determine syllable onsets as maximal positive peaks in the velocity of low-pass filtered amplitude envelope. In order to refine the quality of the segmentation, the algorithm ensures that the detected peaks are followed by a potential nucleus, a sonorant sound with prominent energy below 1-kHz (F1-region), and also inhibits smaller peaks within ±100-ms from prominent peaks from being considered as segment boundaries (see [26] for details).

In the original paper VSeg was shown to compare favorably against the classical convex-hull algorithm [28] and has since been used in, e.g., paralinguistic speech processing [24]. In the present work, we use a MATLAB implementation of the algorithm with the AuToBI-toolbox modifications [27, 29]. After manual experimentation, we ended up using the same default parameters for the algorithm as described in the original paper [26] and also used in the AuToBI.

2.1.2. Envelope minima detector

Although not necessarily optimal with respect to linguistic definition of syllables, direct minima detection from smoothed amplitude envelopes has the potential to yield rhythm-based segmentation of speech without the risk of overfitting algorithm parameters to specific languages or speaking styles (see also [30] for a comparison to other methods).

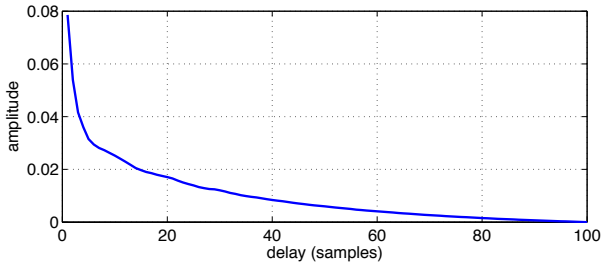

Figure 2: *Impulse response of the low-pass filter used to compute amplitude envelope of speech (f_s = 1000 Hz).*

In the current minima-based algorithm, we first take the absolute value of the incoming signal and then downsample the signal to 1000 Hz. We smooth the waveform in time using a 100-point FIR filter (Fig. 2) that approximates the shape of the temporal window of integration in human hearing (see [17]). In comparison to the standard moving average filter, this minimum-phase filter maintains sharp onset-detection while still providing efficient smoothing of the envelope with approximately 7-Hz cutoff-frequency.

In order to detect syllable boundaries, the envelopes are normalized to a value range of [0, 1] and then each local minimum that is preceded or followed by a local maximum of at least δ units higher is marked as a candidate boundary. Boundaries closer than 50-ms to each other are replaced with a single boundary located at the deeper envelope trough of the two candidates. In the present experiments, $\delta = 0.12$ was used, as this led to a maximal number of segments with durations between 200- and 500-ms on American English speech.

2.1.3. Amplitude envelope-driven oscillator

The third segmentation algorithm was inspired by neurophysiological models of speech perception [20, 21] where theta-rate oscillations, coupled to the speech envelope, are assumed to be responsible for providing timing to the speech perception (see Section 1.1). Therefore, we used a simple damped harmonic oscillator, driven by the amplitude envelope of speech, as a model of auditory entrainment to the syllabic rhythm.

The envelope that was used to drive the oscillator was computed similarly to the envelope of the minima-detection algorithm by downsampling full-wave rectified waveforms to 1000 Hz and then low-pass filtering them with the filter in Fig. 2. The oscillator's behavior in a discrete-time system was modeled using the following equations:

$$\begin{aligned} f(t) &= e(t) - kx(t-1) - cv(t-1) \\ v(t) &= v(t-1) f(t) / (f_s m) \\ x(t) &= x(t-1) v(t) / f_s \end{aligned} \quad (1)$$

where $f(t)$, $a(t)$, $v(t)$ and $x(t)$ are the force, acceleration, velocity, and amplitude of the oscillator at time t, respectively. Also, $e(t)$ is the speech amplitude envelope, c is the damping coefficient, k is the spring constant, and m is the mass of the oscillator. According to the physics of harmonic oscillation, the spring constant can be fixed to $k = 1$ and the mass of an oscillator with a desired center frequency f_0 is then obtained by

$$m = 1 / (4\pi^2 f_0^2) \quad (2)$$

The damping coefficient c leading to a desired bandwidth Δf is then obtained by

$$c = (\Delta f \sqrt{m}) / f_0 \quad . \quad (3)$$

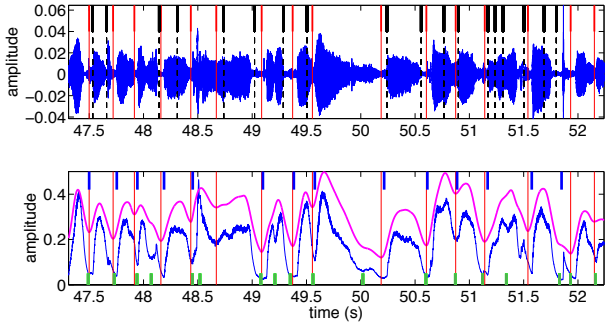

Figure 3: *An example of segmentation with the oscillator. Top panel: Original waveform. Bottom panel: Amplitude envelope (the blue line) and oscillator amplitude (the magenta line). Detected boundaries (oscillator minima) and reference word boundaries are shown with vertical solid red and dashed black lines, respectively. VSeg and EnvMin boundaries are shown in the bottom panel with short green and blue lines, respectively.*

To roughly match the oscillator to the syllabic rhythm of speech (and thereby to theta-rhythm of brain oscillations), the center frequency was set to $f_0 = 4$ Hz and bandwidth to $\Delta f = 8$ Hz (critical damping). The segmentation was carried out by feeding the speech envelope to the oscillator and marking all oscillator minima as segment boundaries. Phase-shift (approx. 70 ms) between the envelope and the oscillator amplitude was compensated automatically by finding a constant delay that minimized the RMSE between the envelope and the oscillator amplitude across the entire signal. Fig. 3 shows an example of the segmentation process.

2.2. Feature extraction and clustering

Following the conceptual models of Ghitza [20] and Giraud & Poeppel [21], we assume that the syllabic rhythm provides frames (or "information packages") within which more rapid sampling of detailed signal content takes place.

To describe the spectral content of each syllable, standard MFCCs were used. More specifically, the first 12 MFCC coefficients and energy were first computed for the signals using a 25-ms window size and 10-ms step size, followed by cepstral mean and variance normalization across the recording. Then each discovered syllable segment i was uniformly divided into N disjoint sub-segments in time and the mean of the MFCC vectors $\mathbf{y}_{i,j}$ falling within each sub-segment j were computed. Finally, the sub-segment MFCCs were concatenated into one fixed-length feature vector together with a scaled log-duration d_i of the syllable:

$$\mathbf{y}_{i,\text{tot}} = [\mathbf{y}_{i,1}{}^{\text{T}}, \mathbf{y}_{i,2}{}^{\text{T}}, \ldots, \mathbf{y}_{i,N}{}^{\text{T}}\ N/3 * \log(d_i)]^{\text{T}} \quad (4)$$

The scaling factor $N/3$ was set empirically to balance the scale of duration with the spectral content of the syllable tokens.

Instead of using uniform temporal division, we also experimented with a faster (20–40 Hz) oscillator coupled to the syllabic-oscillator or to the envelopes of a Gammatone-filterbank in order to segment syllables into sub-syllabic segments. Since both approaches led to very similar results as those obtained with uniform slicing of the syllables, the current results are reported using the simplest uniform segmentation with the number of sub-segments set to $N = 5$.

In order to find recurring syllables, the syllable feature vectors were clustered in an unsupervised manner into Q clusters using the standard k-means algorithm. Clustering was carried out separately for each talker and the process was initialized by randomly sampling from the full set of syllable tokens from the talker. Speaker-specific clustering was chosen because the acoustic variability in the present material was too high to achieve notable improvements in performance by pooling patterns across multiple talkers, even after unsupervised vocal tract length normalization. We also investigated agglomerative clustering of syllable tokens using DTW and observed very similar results to the uniform spectral slicing approach. We also replicated this finding on Brent corpus [31] of infant-directed speech. This suggests that the entrainment to syllabic rhythm provides automatic temporal normalization for speech patterns and therefore separate time-alignment is not needed for pattern matching purposes.

In the present experiments, the number of clusters was set to 30% of the overall number of syllable tokens for a given talker. This parameter was set to balance the set of frequently recurring syllables with the large set of syllable types that occurred only once in a given talker's data. We also tried to set the number of clusters to the expected number of unique syllable types based on Zipf's law, but we found that estimate too low to account for the acoustic variability in the data.

2.3. Word decoding with n-grams

After clustering, monosyllabic words are in principle already represented as clusters. In order to discover multisyllabic words, we applied standard n-gram modeling to find recurring sequences of syllables. We started from the longest recurring n-grams ($n = 3$ in practice) and found all n-grams of that order that occurred at least twice in the data. Syllables that were part of these n-grams were excluded from further analysis and the process was repeated for the n-grams of the next highest order. The process was done all the way to unigrams by including all remaining unigrams as patterns. The output of the process was a list of pattern locations and their corresponding identifiers.

3. Experiments

3.1. Data and evaluation

Evaluation of the system was performed on the Zerospeech-challenge data sets. The data consist of two different corpora: one of conversational speech in American English, the Buckeye corpus [32], and a corpus of Tsonga speech [33]. As defined by the challenge, a 10.5 h subset of the Buckeye corpus was used for training. A total of 12 unique talkers contributed English data; all speech were recorded during interview sessions with a head-mounted microphone in a seminar room. Tsonga data were recorded in the field using the Woefzela mobile phone data collection app [33]; this dataset contained a total of 4.4 hours of speech from 24 different talkers. Both datasets were provided with evaluation intervals that specified the timestamps for speech by the talkers of interest and excluded periods of silence or overlapping speech from another talker [25].

3.2. Evaluation metrics

All evaluations were performed using the Zerospeech evaluation kit described in [34]; the reader is directed to the original paper for full technical details. The basic method in the kit is to represent each discovered pattern as a sequence of phonemes of which at least 50% or 30-ms are covered by the pattern. Two basic aspects of the learned patterns are then measured, 1) the normalized edit distance ("*NED*") between

Table 1: *Results for the Zerospeech task using the three different segmentation algorithms. All measures are reported in percentages.*

	general		phoneme grouping			word token			word type			word boundary		
English	NED	cov	PRC	RCL	F	PRC	RCL	F	PRC	RCL	F	PRC	RCL	F
Baseline	**21.9**	16.3	**21.4**	**84.6**	**33.3**	5.5	0.4	0.8	6.2	1.9	2.9	44.1	4.7	8.6
VSeg	89.6	40.6	4.0	10.8	5.7	21.6	4.8	7.9	13.5	11.3	12.3	**76.1**	28.5	41.4
EnvMin	88.0	42.2	4.3	10.6	6.0	21.6	4.7	7.8	12.7	10.8	11.6	75.7	27.4	40.3
Osc	70.8	**42.4**	13.4	15.7	14.2	**22.6**	**6.1**	**9.6**	**14.1**	**12.9**	**13.5**	75.7	**33.7**	**46.7**
Tsonga	NED	cov	PRC	RCL	F	PRC	RCL	F	PRC	RCL	F	PRC	RCL	F
Baseline	**12.0**	16.2	**52.1**	**77.4**	**62.2**	**2.6**	0.5	0.8	3.2	1.4	2.0	22.3	5.6	8.9
VSeg	78.4	77.7	6.2	3.2	4.2	1.8	1.8	1.8	1.7	4.1	2.4	26.2	26.3	26.3
EnvMin	61.2	**95.2**	17.7	2.8	4.9	0.8	1.3	1.0	1.1	3.3	1.7	16.3	24.4	19.5
Osc	63.1	94.7	10.7	3.3	5.0	2.3	**3.4**	**2.7**	2.2	**6.2**	**3.3**	**29.2**	**39.4**	**33.5**

all phoneme sequences belonging to the same pattern class (cluster), and 2) the proportion of the corpus covered by the learned patterns ("*cov*"). In addition, a more detailed "grouping" analysis is carried out by comparing the consistency of phoneme sequences within each class in terms of precision (how selective each class is to the most dominant phoneme-sequence in the class) and recall (how large a proportion of the same phoneme sequences in all data is covered by the given class), and their harmonic mean, F-score. In addition, evaluation is performed using standard NLP precision and recall measures at the word-level, i.e., how well word tokens and types are captured by the discovered classes. Finally, the accuracy of word segmentation is measured, describing how accurately the boundaries of discovered phoneme sequences correspond to the actual word boundaries. Note that the "matching"-measure of the toolkit was not included as the current system does not attempt pairwise matching of patterns.

3.3. Results

The results for all three segmentation algorithms and both languages can be seen in Table 1. Challenge baselines using the JHU system [5] with PLP-features are also shown.

The most prominent finding is that the syllabic approach leads to high word segmentation accuracy on both languages in comparison to the baseline system. In English, 75% of the hypothesized boundaries match a true word boundary with the accuracy of one phoneme while the found patterns still cover more than 40% of the speech data. This also results in significantly higher word token and type accuracies than those reported with the baseline system. In contrast, the baseline system is superior in the consistency of the phoneme sequences it is finding, but is also much more selective, covering only 16% of the English data and finding only 5% of the word boundaries. Manual listening to the syllabic patterns revealed that there are a number of very pure clusters containing repetitions of one word type or filler type, and a large number of more mixed clusters containing multiple different syllables and/or words. Pruning of clusters with largest acoustic distortion between tokens led to improvements in NED, but at the cost of notable decreases in recall measures.

As for the syllabification algorithms, it appears that the oscillator-based approach outperforms the other algorithms with a clear margin both in terms of segmentation accuracy and in terms of token/type accuracy on both languages.

Cross-linguistically, the token/type performance is much higher on English than on Tsonga while Tsonga-clusters are more pure on average. This stems from the higher rate of multisyllabic words in Tsonga. Since there is limited amount of data from each Tsonga talker (on average 11 min), the algorithm is not capable of finding these syllabic sequences in a reliable manner due the to inaccuracies at the clustering stage. In practice, no recurring 4-grams or higher order n-grams exist in the cluster sequences.

4. Discussion and conclusions

The current results show that the speech amplitude envelope contains strong cues for word segmentation in purely unsupervised settings, enabling syllabification and thereby word boundary discovery with a small number of a priori assumptions. It was also shown that a simple oscillator model performs well in the syllabification task, providing interesting parallels to the brain research on human speech perception.

An obvious drawback in using the syllable-segmentation is that the success in the later pattern matching stages is critically dependent on the consistency of the initial segment boundaries. At the temporal scale of syllables, any insertions or deletions in boundaries are destructive to the compatibility of the resulting feature representations unless there are additional mechanisms to handle this uncertainty. In addition, it appears that Euclidean distance-based clustering in the MFCC space is not sufficient for discriminating between syllable classes while still capturing majority of the same-class tokens. In the present system, this translates to to a high coverage but also to a relatively high variety of acoustic contents in many of the clusters. In the future, this issue should be addressed by improving clustering at the syllable token level or by intelligent pruning of the initial clusters based on their contents. Utilization of information at multiple levels of representation could also be studied in order to refine the initial syllable classes (see, e.g., [35,36]).

As a final remark, the overall machine time for processing the 10.5 h English data was approximately 20 minutes, translating to 30x faster than real-time-processing with a non-optimized MATLAB implementation. This makes the syllabic approach a potential pre-processing step for more advanced pattern discovery and matching algorithms such as the previously reported methods, and is also suitable for on-line systems with limited computational resources.

5. Acknowledgements

This research was funded by the Academy of Finland and by ONR grant N00014-13-1-0287. We also acknowledge the computational resources provided by the Aalto Science-IT project used for running the Zerospeech evaluation software. The MATLAB codes of the algorithms will be made available for download at: http://users.spa.aalto.fi/orasanen/zerospeech.

6. References

[1] J.R. Glass, "Towards unsupervised speech processing," *Proc. ISSPA-2012*, Montreal, QC, pp. 1–4, 2012.

[2] P.K. Kuhl, "Early language acquisition: cracking the speech code," *Nature Reviews Neuroscience*, vol. 5, pp. 831–843, 2004.

[3] A.S. Park and J.R. Glass "Unsupervised Pattern Discovery in Speech," *Proc. ICASSP'08*, Las Vegas, Nevada, pp. 186–197, 2008.

[4] Y. Zhang and J. Glass, "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriograms," *Proc. ASRU-2009*, Merano, Italy, pp. 398-403, 2009.

[5] A. Jansen and B. Van Durme, "Efficient spoken term discovery using randomized algorithms," *Proc. IEEE ASRU Workshop,* Waikoloa, Hawaii, pp. 401–406, 2011.

[6] A. Muscariello, G. Gravier, and F. Bimbot, "Unsupervised Motif Acquisition in Speech via Seeded Discovery and Template Matching Combination," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, no.7, pp. 2031–2044, 2012.

[7] O. Räsänen, "A computational model of word segmentation from continuous speech using transitional probabilities of atomic acoustic events," *Cognition*, vol. 120, 149–176, 2011.

[8] M-h. Siu, H. Gish, A. Chan, W. Belfield, and S. Lowe, "Unsupervised training of an HMM-based self-organizing unit recognizer with applications to topic classification and keyword discovery," *Computer Speech and Language*, vol. 28, pp. 210–223, 2014.

[9] T. Oates, "PERUSE: An unsupervised algorithm for finding recurrent patterns in time-series," *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Maebashi City, Japan, pp. 330–337, 2002.

[10] A. Cutler and D. Norris, "The role of strong syllables in segmentation for lexical access," *Journal of Experimental Psychology*, vol. 14, pp. 113–121, 1988.

[11] P.W. Jusczyk and D.M. Houston, "The beginnings of word segmentation in English-learning infants," *Cognitive Psychology*, vol. 39, pp. 159–207, 1999.

[12] P. Eimas, "Segmental and syllabic representations in the perception of speech by young infants," *J. Acoust. Soc. Am.*, vol. 105, pp. 1901–1911, 1999.

[13] E. Dupoux, "The time course of prelexical processing: The syllabic hypothesis revisited," In G. and S. Altmann (Eds.) *Cognitive Models of Speech Processing*, (pp 81-114) Hillsdale, NJ: Erlbaum, 1993.

[14] V. Leong, M. Kalashikova, D. Burnham, and U. Goswani, "Infant-directed speech enhances temporal rhythmic structure in the envelope," *Proc. Interspeech'2014,* Singapore, 2014.

[15] R. Plomp and M.A. Bouman, "Relation between hearing threshold and duration for tone pulses," *J. Acoust. Soc. Am*., vol. 31, pp. 749–758, 1959.

[16] N.F. Viemeister and G.H. Wakefield, "Temporal integration and multiple looks," *J. Acoust. Soc. Am*., vol. 90, pp. 858–865, 1991.

[17] O. Räsänen and U.K. Laine, "Time-frequency integration characteristics of hearing are optimized for perception of speech-like acoustic patterns," *Journal of the Acoustical Society of America*, vol. 134, pp. 407–419, 2013.

[18] J. Mehler and R. Hayes, "The role of syllables in speech processing: Infant and adult data," *Philosophical Transactions of the Royal Sociecty of London, Series B, Biological Sciences,* vol. 295, pp. 333–352, 1981.

[19] E. Ahissar, S. Nagarajan, M. Ahissar, A. Protopapas, H. Mahncke, and M. Merzenich, "Speech comprehension is correlated with temporal response patterns recorded from auditory cortex," *PNAS*, vol. 98, pp. 11367–13372, 2001.

[20] O. Ghitza, "Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input," *Frontiers in Psychology*, vol. 2, pp. 1–13, 2011.

[21] A-L. Giraud and D. Poeppel. "Cortical oscillations and speech processing: emerging computational principles and operations," *Nature Neuroscience*, vol. 15, pp. 511–517, 2012.

[22] M. J. Hunt, M. Lennig, and P. Mermelstein, "Experiments in syllable-based recognition of continuous speech," *Proc. ICASSP-1980*, Denver, Colorado, pp. 880–883, 1980.

[23] A. Ganapathiraju, J. Hamaker, J. Picone, M. Ordowski, and G.R. Doddington," Syllable-based large vocabulary continuous speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 9, pp. 358–366, 2001.

[24] G. An, D. Brizan, and A. Rosenberg, "Detecting laughter and filler pauses using syllable-based features," *Proc. Interspeech'2013*, Lyon, France, pp. 178–181, 2013.

[25] M. Versteegh, R. Thiolliere, T. Schatz, X.N. Cao, X. Anguera, A. Jansen, and E. Dupoux, "The zero resource speech challenge," *Proc. Interspeech-2015*, Dresden, Germany, 2015.

[26] R. Villing, J. Timoney, T. Ward, and J. Costello, "Automatic Blind Syllable Segmentation for Continuous Speech," *Proc. Irish Signals and Systems Conference (ISSC 2004)*, Belfast, Northern Ireland, 2004.

[27] A. Rosenberg. 2010. AuToBI - A Tool for Automatic ToBI Annotation. *Proc. Interspeech-2010,* Makuhari, Japan, pp. 146–149, 2010.

[28] Mermelstein, P., "Automatic Segmentation of Speech into Syllabic Units", *J. Acoust. Soc. Am.*, Vol. 58, No. 4, 880–883, 1975.

[29] https://github.com/vsoto/autobi/blob/master/src/edu/cuny/qc/speech/AuToBI/Syllabifier.java. Accessed March 4th, 2015.

[30] R. Villing, T. Ward, and J. Timoney, "Performance limits for envelope-based automatic syllable segmentation," *Proc. ISSC-2006*, Dublin, Ireland, pp. 521–526, 2006.

[31] M.R. Brent & T.A. Cartwright, "Distributional regularity and phonotactic constraints are useful for segmentation", *Cognition*, vol. 61, pp. 93–125, 1996.

[32] M.A. Pit, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier, "Buckeye Corpus of Conversational Speech (2nd release)," Columbus, OH: Department of Psychology, Ohio State University (Distributor), 2007.

[33] N.J. de Vries et al., "A smartphone-based ASR data collection tool for under-resourced languages," *Speech Communication*, vol. 56, pp. 119–131, 2014.

[34] B. Ludusan, M. Versteegh, A. Jansen, G. Gravier, X.N. Cao, M. Johnson, and E. Dupoux, "Bridging the gap between speech technology and natural language processing: an evaluation toolbox for term discovery systems," *Proc. LREC-2014*, Reykjavik, Iceland, 2014.

[35] A. Fourtassi and E. Dupoux, "A rudimentary lexicon and semantics help bootstrap phoneme acquisition," *Proc. CoNLL-2014*, Ann Arbor, Michigan, pp. 191–200, 2014.

[36] S. Frank, N. Feldman, and S. Goldwater, "Weak semantic context helps phonetic learning in a model of infant language acquisition," *Proc. 52nd Annual Meeting of the Association of Computational Linguistics*, Batimore, Maryland, pp. 1073–1083, 2014.