Basic cuts revisited: Temporal segmentation of speech into phone-like units with statistical learning at a pre-linguistic level

Okko Räsänen (okko.rasanen@aalto.fi)
Department of Signal Processing and Acoustics, Aalto University
PO Box 13000, 00076 Aalto, FINLAND

Abstract

Considerable effort has been put to understand how infants may utilize statistical regularities of speech in early word segmentation. Some studies suggest that infants are able to discover word boundaries at the points of high unpredictability across subsequent linguistic units such as phonemes or syllables. Meanwhile, the possible role of the statistical regularities in the temporal organization of the speech at a pre-linguistic acoustic level has not been widely addressed. The current work examines how the short-term temporal predictability of the acoustic speech signal correlates with linguistically motivated phone-, syllable-, and word-level units. The results indicate that the points of low predictability correlate mainly with the boundaries between phone-like segments. This suggests that the same statistical learning mechanisms hypothesized to operate at the word level can also aid in temporal organization of the speech stream into phone-like temporal segments before knowing the phonemic or syllabic units of the language.

Keywords: distributional learning; language acquisition; phone segmentation; speech segmentation; statistical learning

Introduction

Segmentation of continuous speech into linguistically relevant units is essential for successful language acquisition (LA). Segmentation can take place at a number of levels, as the speech can be linguistically characterized in terms of units such as phones, syllables, and words, and with the latter always consisting of the former.

In the early LA research, infants' ability to segment words from speech has received a large amount of attention as the words are the main functional units of the language, standing for entities, events, actions, and states of the surrounding world. In the word segmentation studies, one of the major findings is that the infants can use statistical regularities in the speech input in order to discover boundaries between words (Saffran, Aslin & Newport, 1996). Also, these statistical learning mechanisms do not seem to be specific to words or even language faculty but operate across many levels of representation and perceptual domains (see, e.g., Romberg & Saffran, 2010, for a recent review).

Importantly, a large body of the existing work on statistical word learning assumes that the infants are capable of representing speech input in terms of linguistically relevant units such as phones or syllables. Given the representational units, the infants are supposedly tracking transitional probabilities (TPs) between these units across time and use low-probability transitions as indications for word boundaries while the high-probability regions form representational units (Saffran et al., 1996). This strategy is valid as long as the TPs within words are higher than the TPs across word boundaries. However, the infant's access to linguistic units such as phones or syllables and their statistics cannot be taken for granted. It is still unclear whether early adaptation to phonetic units drives lexical learning (c.f., NLM-e theory by Kuhl et al., 2008) or whether early lexical learning actually precedes, or at least parallels, the acquisition of sub-word representation of spoken language (e.g., Werker & Curtin, 2005). The "sub-word units –first" approach is challenged by the fact that the bottom-up organization of speech signal into temporally and categorically discrete units is far from trivial. Learning a phonetic or syllabic representation of the spoken language includes both the *segmentation problem* (division of the signal in time) and the *categorization problem* (assigning context-, talker-, and speaking style-dependent acoustic observations into a correct number of linguistic categories). Importantly, infants do not have access to any ground truth in either of the two tasks while learning the native language, suggesting that some speech-external factors such as feedback from lexical level or social interaction are required for successful learning.

Still, it seems that even the basic problem of segmenting speech into sub-word units has been largely overlooked in the existing LA research. For example, it is unclear how well natural co-articulated speech can be segmented into sub-word units before learning the phonetic or lexical units of the language, and whether infants actually do such segmentation. Possibly the most concrete reference to early sub-word segmentation in the existing literature is the Kuhl's concept of *basic cuts*: a perceptual mechanism that provides an initial low-level chunking of the speech stream into primitive phone-like units and which then gradually improves towards native language phone system through language exposure (Kuhl, 2004, and references therein). Segmentation into syllabic units is also central to many theories of LA (e.g., Jusczyk, 1993) although explicit and well-controlled studies on the segmentation process itself are few.

In the speech engineering community, both phone- and syllable-level segmentation have been widely studied. The general finding is that the spectral changes (or "jumps") in speech are good candidates for phone boundaries as they correlate with the changes in articulator positions (e.g., Almpanidis & Kotropulos, 2008; Esposito & Aversano, 2005; ten Bosch & Cranen, 2007; Scharenborg et al., 2007). On the other hand, it is known that syllabic segmentation

can be achieved by detecting minima from the smoothed temporal envelope of speech signals (see Villing, Ward & Timoney, 2006, for a performance overview). It is likely that the auditory system achieves "basic cuts" based on an innate perceptual mechanism that detects sufficiently large spectral changes in the input and/or uses the temporal envelope to parse speech into rhythmic units.

However, there is another open possibility that it is not the magnitude of the spectral or envelope change as such that drives the segmentation processes, but maybe the *short-term statistical regularities* of the acoustic speech signal enables segmentation of the input into perceptually relevant units. As it is already known that the distributional learning plays a role in the word segmentation (Saffran et al., 1996) and in the categorization of native speech sounds (e.g., Maye, Werker & Gerken, 2002; Kuhl, 2004), it is of interest whether similar learning mechanism could aid the organization of the speech into syllabic or phonetic units in time. If this would be the case, then only a single learning mechanism operating on different levels of representation would be needed to explain both early low-level sub-word organization, word-level segmentation (Saffran et al., 1996), and many other aspect of perceptual processing associated with statistical learning (see Romberg & Saffran, 2010 and references therein).

In order to investigate the sub-word segmentation from statistical learning point of view, the current paper presents results from simulations where the transition probability analysis is carried out at the level of millisecond-scale acoustic features. The hypothesis is that the points of low TP in time have some correspondence to the boundaries between linguistically motivated units, and therefore we compare the model output to manual transcription of the signals at the phone-, syllable-, and word-levels.

Data

TIMIT corpus (Garofolo et al., 1993) containing American English continuous speech from multiple talkers and dialects was chosen for the experiments due to its rich and balanced phonetic content and due to the availability of high-quality phone- and word-level transcriptions of the utterances. Since TIMIT is recorded in a controlled noise-free environment, the focus is purely on the analysis of speech structure without any interfering effects from background noise or, e.g., multiple overlapping talkers.

As the original TIMIT only contains phone- and word-level transcriptions, syllable annotation was generated from the phonetic transcription using the tsylb2-algorithm (Fisher, 1996) that uses the phonological rules described in Kahn (1976) for the transformation. Phonetic alphabet used in tsylb2 was matched to the TIMIT in a similar fashion to the study of Villing, Ward & Timoney (2006). The syllabic transformation was carried out using the tsylb2 parameters associated with "*ordinary conversational speech*". The phone level boundaries were used as they are described in the original TIMIT format. This includes the boundaries between plosive closures and bursts (e.g., [k] + [kcl]) since they can be considered as articulatory distinct segments.

In the simulations, the standard TIMIT NIST training set (462 talkers, 4620 utterances, both male and female talkers) was used to learn the TPs between the acoustic events (see Methods). Then the NIST core test set containing 192 previously unseen utterances from 24 talkers was used to evaluate the segmentation performance. Overall duration of the data was approx. 4 hours (177080 phone segments) for training and 10 minutes (7333 phones) for testing.

Methods

The basic acoustic unit analyzed in the current work consists of spectral features that are computed from fixed-size short-term (millisecond scale) segments of speech. These features are then quantized into Q possible signal *states* in an unsupervised manner and TPs between the states are used as a model for acoustic predictability of the speech (Figure 1). Finally, points of low TP are extracted as candidate segment boundaries. As the TP analysis is carried out in an abstract state space, the model is agnostic to the exact magnitude of the spectral changes but the changes are simply reflected in the state changes across time.

Importantly, the obtained signal states do not correspond to phonetic categories of the language as the bottom-up clustering of spectral features into talker- and context-independent phonemic units is not possible without additional information such as lexical knowledge or articulatory constraints (e.g., Feldman, Griffiths & Morgan, 2009; see Räsänen, 2012, for a review). The quantization simply acts as a conversion from the continuous multivariate input into a discrete categorical sequence suitable for standard TP analysis. However, the clustering used to create the quantization codebook will necessarily introduce a rough "perceptual re-organization by language exposure" as the cluster boundaries will reflect the distributional characteristics of the speech spectra.

Note that the current work does not imply that infants would analyze acoustic signal in terms of Q different discrete units or categories (as it is unlikely that infant brain would represent a discrete probability distribution for TPs between discrete syllables; cf., Saffran et al., 1996). Instead, the goal of the pre-processing and quantization is to simply enable the analysis of statistical regularities in the signal using the simplest possible mathematical form similarly to the discussion on "tracking of TPs" in the context of perceptual learning.

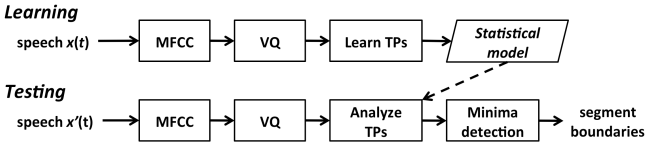

Figure 1: A schematic view of the TP-based segmentation process. VQ stands for vector quantization.

Pre-processing of speech

One of the challenges with the acoustic analysis is that the relevant units are not known in advance. This means that the raw speech signal has to be represented using non-linguistic features that capture similar time-frequency information than what the auditory system is capable of extracting. Here, standard Mel-frequency cepstral coefficients (MFCCs) were used as they compactly represent the essential spectral content of speech with low-dimensional feature vectors and approximate the spectral resolution of human hearing.

MFCCs are obtained by first computing the power spectrum of the speech signal using fast Fourier transform (FFT) in a sliding window of length 20 ms and a step size of 10 ms. For each window position, the obtained FFT-spectrum is filtered through a Mel-scale filterbank with 26 triangular bandpass filters in order to approximate the frequency resolution of the auditory system. Finally, the logarithm of the Mel-spectrum is taken and discrete cosine transform is applied to the log-Mel spectrum to obtain the MFCC coefficients (Figure 2, second panel). The first 12 coefficients $c_1 \ldots c_{12}$ and the c_0 coefficient corresponding to the signal energy were chosen for further processing as they are sufficient for describing the spectral envelope of speech. Mean and variance of each cepstral coefficient was z-score normalized across each utterance before further processing.

In order to perform TP analysis on the spectrum, MFCCs were quantized into a discrete state space by first clustering 10000 randomly chosen MFCC vectors of the training data into a codebook of Q clusters with the standard k-means algorithm. Then all MFCC vectors were assigned to the nearest cluster centroid in terms of Euclidean distance and replaced by the corresponding state index. As a result, the speech signal of L frames becomes represented as a sequence of discrete states $X = \{w_1, w_2, \ldots, w_L\}$, $w \in [1, Q]$, $t \in [0, L]$, with one state occurring every 10 ms (Figure 2, third panel; see also Räsänen, 2011).

Transition probability analysis

During training, the TPs between subsequent states were computed for a number of lags $k = \{1, 2, 3, \ldots, K\}$, where a lag k transition means a state-pair $\{w_{t-k}, w_t\}$ with any undefined elements $w_{t-k+1} \ldots w_{t-1}$ in between. The probability of the signal X as function of time was defined as

$$p(t - \lfloor k/2 \rfloor \mid X) = \sum_{k=1}^{K} p(w_t \mid w_{t-k}) \qquad (1)$$

where $\lfloor$ denotes downward rounding to an integer. The $p(w_t|w_{t-k})$ were simply calculated from the transition frequencies $f(w_t|w_{t-k})/f(w_{t-k})$ counted from the training data. As defined in Eq. (1), the statistics of the signal were modeled as a mixture of TPs at different temporal distances across the current time frame of analysis, corresponding to an approximation of a higher-order Markov chain but making it learnable from finite data. This allowed the model to capture the temporal dependencies that extend beyond the neighboring states as the acoustic dependencies in speech are known to extend up to approximately 250-ms in time

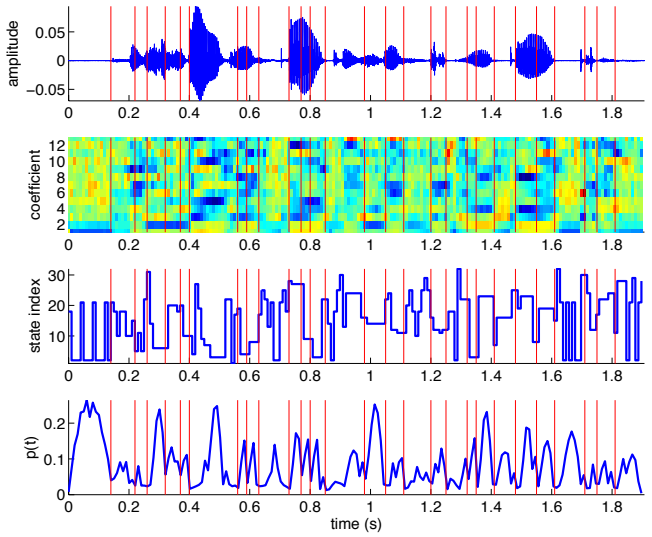

Figure 2: An example of the processing stages for an utterance "*His shoulder felt as if it were broken*". Top: The original speech waveform. Second panel: MFCC spectrum computed from the speech signal. Third panel: Corresponding VQ-state indices. Bottom: Transition probability (TP) curve. Red vertical lines show the minima, a.k.a. the boundary hypotheses, extracted from the TPs.

and that the human auditory system also analyzes signal content on the same time scale (Räsänen & Laine, 2013).

During the segmentation stage, probabilities of the transitions in a previously unseen signal $X' = \{w_1, w_2, \ldots\}$ were simply evaluated according to Eq. (1), leading to a probability curve as a function of time (Figure 2, bottom). The final set of low probability points (LPPs) were extracted from the probabilities by using a simple valley detection procedure. A segment boundary was hypothesized to each local minimum that was preceded by a TP-value larger by at least δ units, where δ is a user set parameter. The use of a fixed global threshold for minima detection was also studied and it was found to lead to very similar results than the local minima detection procedure. However, the fixed threshold requires additional rules to deal with multiple neighboring points that are all below the threshold in order to avoid unnecessary over-segmentation.

Evaluation

The overall segmentation quality was evaluated in terms of the overall agreement between the LPPs and the reference annotation, quantified by the F-value in Eq. (2) that is obtained as the harmonic mean of the precision in Eq. (3) and recall in Eq. (4).

$$\mathrm{F} = 2 * \mathrm{PRC} * \mathrm{RCL}/(\mathrm{PRC} + \mathrm{RCL}) \qquad (2)$$

$$\mathrm{PRC} = N_{\mathrm{hit}}/N_{\mathrm{hypo}} \qquad (3)$$

$$\mathrm{RCL} = N_{\mathrm{hit}}/N_{\mathrm{ref}} \qquad (4)$$

In the equations, N_{hit} is the number of correctly detected segment boundaries, N_{hypo} is the total number of boundary

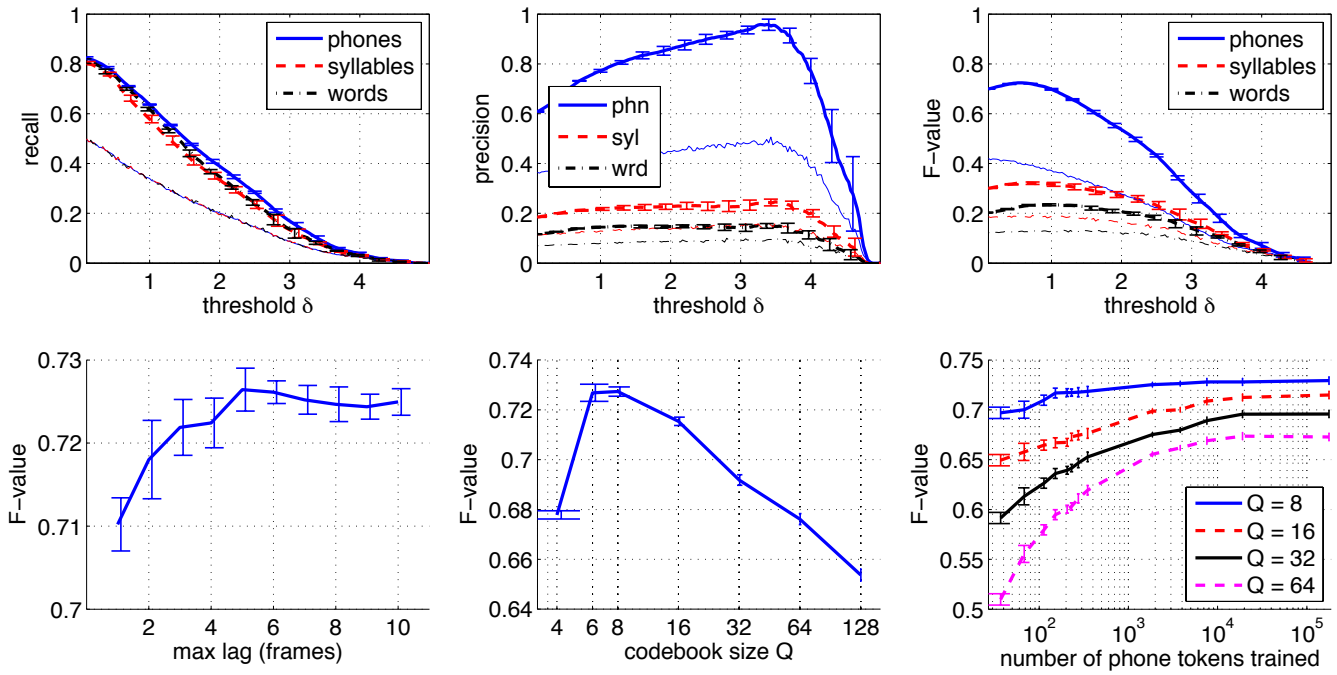

Figure 3: Results from the simulations. **Top row:** phone, syllable, and word segmentation results with the best parameter combination of $K = 10$ and $Q = 8$ as a function of the detection threshold δ. Y-axis shows the recall (left), precision (center), and F-value (right). The thick blue solid lines, the red dashed lines, and the black dash-dotted lines correspond to the results calculated with respect to the annotated phone, syllable, and word boundaries, respectively. Thin lines show the corresponding baseline performances from the random boundary generation. Standard deviations (SDs) across multiple runs are shown with horizontal bars. SDs of the random baselines are not shown for the sake of visual clarity but are of the same scale as the SDs of the other results. **Bottom row**: Phone segmentation F-value as a function of the maximum number of lags K in Eq. (1) with fixed $Q = 8$ (left), F-value as function of the quantization codebook size with fixed $K = 10$ (center), and F-value as a function of the training data length (measured in phone tokens) for different codebook sizes (right).

hypotheses generated by the model, and N_{ref} is the total number of reference boundaries in the annotation.

For a reference phone boundary to be considered as correctly detected, the algorithm was required to produce a hypothesized boundary within ±20 ms of the reference boundary as this roughly corresponds to the variability in the phonetic annotation across multiple annotators (Kvale, 1993). Since the syllable- and word boundaries are a subset of the phone boundaries in the annotation, the allowed deviation for syllables and words was also set to ±20 ms.

Chance-level performance was measured for all test conditions by generating the same number of boundaries for each utterance than what was produced by the actual algorithm and randomizing the final locations of the boundaries along the utterance duration.

Results

Figure 3 shows the results from the TIMIT core test set segmentation. Top row shows the performance as a function of the detection threshold δ using quantization codebook size of $Q = 8$ and a maximum TP analysis lag of $K = 10$ (100 ms). In the plots, the variability and the associated SDs in the results are caused by the random initialization in the generation of the quantization codebook.

The main observation is that the short-term acoustic dependencies are mainly associated with phone-level structure, TP minima detection leading to notably above chance-level phone segmentation accuracy. In contrast, the syllable- and word-level performances are much worse. Recall for all three levels of representation is approximately equal for all thresholds. On the other hand, precision for phones is always superior to syllables, while precision for syllables is always superior to words. This suggests that the syllable boundaries are simply a subset of the detected phone boundaries without any specific threshold level (depth of minima) being more associated with syllabic structure in comparison to the phones.

In overall, the best phone segmentation result is F = 0.73, corresponding to approximately 70% of boundaries correctly detected with a precision of 74%. This is a surprisingly good performance level considering the lack of specially tailored signal processing solutions typically used to fine-tune the phone segmentation performance. As a reference, the typical performances of dedicated phone segmentation algorithms are in the range of 0.74–0.76 for the F-value on the same TIMIT corpus (e.g., Almpanidis & Kotropulos, 2008; Esposito & Aversano, 2005; Scharenborg et al., 2007).

Regarding the statistical significance, it is evident that the mean phone segmentation performance is far above the chance level for the majority of the threshold values. As for the syllables, the performance is significantly above the chance-level ($p < 0.01$) for thresholds $\delta < 2.65$. In the case of word boundaries, the performance is above chance level ($p < 0.01$) for $\delta < 2.71$.

As for the model parameters, there are four main factors that can affect the results: the size of the codebook, MFCC window size and step size, and the maximum lag K up to which TPs are measured. The size of the MFCC window was found be optimal around 12–30 ms with a step size of 10 ms. No qualitative changes in the relative performance of different linguistic units were observed when these two parameters were adjusted. This finding is expected as the speech signal is quasi-stationary within the given time scale and the FFT step in MFCC computation assumes signal stationarity within the analysis window.

Bottom left panel in Figure 3 shows the phone segmentation performance as a function of the maximum temporal lag K up to which TPs were measured. The performance seems to saturate after the maximum lag of K = 6, confirming that there is useful structure beyond neighboring frames. As for the codebook sizes, the best results are obtained with surprisingly small codebooks of size Q = 6 and 8 (Figure 3, bottom middle). However, the performance is relatively good even for the largest codebook sizes tested. The syllable- and word-level performances (not shown) follow similar saturating trend as a function of lags as the phone level but with notable lower overall performance. As for the codebook size, the syllable- or word-level performances do not change significantly when the codebook size is adjusted (also not shown separately). This is in contrast with the phone level where larger codebooks tend to decrease the overall agreement between the algorithm output and the manually annotated reference.

Finally, the right panel at the bottom of Figure 3 shows the F-value as a function of training data length for different codebook sizes. The result shows that the finer the spectral resolution of the model, the more there is improvement with more learning. Interestingly, it seems that only one or two utterances are sufficient for reasonable performance with very small codebooks. This suggests that part of the phone segmentation with small codebooks is achieved due to the spectral change detection, realized as a transition from a state to another. Since there are more transitions from a state to itself than to other states with small codebooks (see also Figure 2), it may be the case that the majority of the non-self transitions have zero probability at an early stage, leading to a segment boundary. Still, there is significant improvement from longer training times even for Q = 8. For larger codebooks, the effect of learning is more evident as the simple state change detection in these cases would lead to large amounts of over-segmentation. In general, these results confirm that the segmentation is not only based on change detection but properly learned TPs are required for the best performance, although the size of the codebook may impose an implicit tradeoff between change detection and statistical segmentation.

Discussion and conclusions

The current work shows that there is clear temporal statistical structure associated with speech that helps segmentation of the input into phone-like units before any linguistic knowledge is acquired. However, the statistical approach does not exceed the traditional spectral change detection in performance, especially when dedicated phone segmentation algorithms are considered. Actually, the spectral "jumps" and unpredictability of the spectrum can be seen as the two sides of a same coin where one always has a consequence to another. Therefore the current study does not argue that the "basic cuts" in the auditory system would be necessarily based on statistical predictability of the signal. Instead, the current work simply shows that there is a probabilistic interpretation to the low-level temporal organization of the speech signal and a simple statistical learning mechanism has the potential to adapt to this structure in order to parse the signal into units that roughly correspond to linguistically defined phones. Note that the statistical learning here refers broadly to the use of recurring similarities in the signal and not to the explicit analysis of TPs between abstract discrete states. Instead, the TP analysis should be seen as a methodological tool to probe the existence or absence of such statistical structure.

Although it is questionable whether a learning-based mechanism to segmentation is more plausible than a simple hard-wired spectral change detector in terms of human auditory processing, the current model is attractive due to its similarity to the behavioral findings on TP-based word-level segmentation (Saffran et al., 1996; see Romberg & Saffran, 2010, for a review) and also to the existing computational models on statistical learning at the acoustic level (see Räsänen, 2012, for a review). For example, if the global TP model in Eq. (1) is partitioned into multiple different models with their own local TP statistics (as in Räsänen, 2011), or gains support from cross-situational visual cues (see Räsänen, 2012), the TP analysis leads to the learning of words instead of phones. Short-term statistical dependencies of speech also explain the how and why the auditory system combines signal input over time in order to form coherent auditory percepts (Räsänen & Laine, 2013), while TP analysis at the level of prosodic features reveals that points of low predictability in these features correlate with perception of stress in speech (Kakouros & Räsänen, accepted for publication). All this evidence suggests that the same basic computational mechanisms operating on signal-level regularities has explanatory power over both sub-word and word level segmentation and on suprasegmental perception of speech. The main difference is only the time-scale of the statistical analysis, acoustic features that are analyzed, and the potential access to additional constraints such as cross-situational cues in other perceptual modalities.

As for the syllable level, it seems that the syllabic segmentation is not straightforward with the spectral features. It seems as if the syllable boundaries are simply a random subset of the phone boundaries in the current simulations. No studied parameter combination (temporal or spectral) was able to provide clear indication of increased precision at the syllable level in comparison to the phone level. However, this is partially expected as the syllabic structure mainly provides a rhythmic frame to the phonetic/phonemic content of speech and is primarily conveyed by the energy envelope of the speech signal, not by the spectral content studied in the current work.

Finally, a note regarding the overall quantitative segmentation performance is in place. Due to the uncertainties associated with the annotation process (see Kvale, 1993), the reference annotation should not be taken as the ultimate ground truth for a perfect division of the speech signal into linguistically defined units. This is even more emphasized in the syllabic reference that is based on a conversion from the phonetic transcription to syllabic units using a set of linguistic rules (Kahn, 1976), not direct annotation of syllabic units based on subjective perception.

In the future work, it would be beneficial to investigate combination of the current model with a statistical model of categorical and lexical learning from real speech. As the quantization of the acoustic input could be gradually improved with distributional learning of the spectral properties related to actual lexical contrasts, this could also lead to improvement in the temporal segmentation. In this way, the entire spectrotemporal parsing of the speech into linguistically relevant units would gradually improve with experience, as already suggested by Kuhl (2004). Also, given a suitable speech corpus, it would be beneficial to replicate the current study using speech from only one or two talkers and infant directed speech to see how the complexity of the data affects the results.

Acknowledgments

This research was funded by the Academy of Finland. The author thanks all reviewers for their very useful comments.

References

Almpanidis, G., & Kotropoulos, C. (2008). Phonemic segmentation using the generalized Gamma distribution and small sample Bayesian information criterion. *Speech Communication*, 50, 38–55.

Esposito, A., & Aversano, G. (2005). Text independent methods for speech segmentation. In G. Chollet et al. (Eds.), *Lecture Notes in Computer Science: Nonlinear Speech Modeling.* Springer Verlag, Berlin, pp. 261–290.

Feldman, N., Griffiths, T., & Morgan, J., (2009). Learning phonetic categories by learning a lexicon. *Proc. 31st Annual Conference of the Cognitive Science Society,* Austin, Texas, pp. 2208–2213.

Fisher, M. W. (1996). tsylb2. National Institute of Standards and Technology. Available online from: http://www.nist.gov/speech/tools.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., & Zue, V. (1993). TIMIT acoustic-phonetic continuous speech corpus. Linguistic Data Consortium, University of Pennsylvania, 1993.

Jusczyk, P. W. (1993). From general to language-specific capacities: the WRAPSA model of how speech perception develops. *Journal of Phonetics*, 21, 3–28.

Kahn, D. (1976). *Syllable based generalizations in English phonology*. Ph.D. dissertation, Department of Linguistics and Philosophy, MIT, Cambridge, 1976.

Kakouros, S. & Räsänen O. (accepted for publication). Statistical Unpredictability of F0 Trajectories as a Cue to Sentence Stress. *Proc. 36th Annual Conference of the Cognitive Science Society*, Quebec, Canada.

Kuhl, P. (2004). Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience*, 5, 831–843.

Kuhl, P., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Phil. Trans. R. Soc. B.,* 363, 797–1000.

Kvale, K. (1993). *Segmentation and Labelling of Speech*. Doctoral Thesis, The Norwegian Institute of Technology, Department of Telecommunications.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82, B101–B111.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Review of Cognitive Science*, 1, 906–914.

Räsänen, O. (2011). A computational model of word segmentation from continuous speech using transitional probabilities of atomic acoustic events. *Cognition*, 120, 149–176.

Räsänen, O. (2012). Computational modeling of phonetic and lexical learning in early language acquisition: existing models and future directions. *Speech Communication*, 54, 975–997.

Räsänen, O., & Laine, U. (2013). Time-frequency integration characteristics of hearing are optimized for perception of speech-like acoustic patterns. *Journal of the Acoustical Society of America*, 134, 407–419.

Saffran, J., Aslin, R. N., & Newport, E. L. (1996). Statistical learning of 8-month-old infants. *Science*, 274, 1926–1928.

Scharenborg, O., Ernestus, M., & Wan, V. (2007). Segmentation of speech: Child's play? *Proc. Interspeech'07*, Antwerp, Belgium, pp. 1953–1956.

ten Bosch, L., & Cranen, B. (2007). A computational model for unsupervised word discovery. *Proc. Interspeech'07,* Antwerp, Belgium, pp. 1481–1484.

Villing, R., Ward, T., & Timoney, J. (2006). Performance limits for envelope based automatic syllable segmentation. *Proceedings of ISSC'2006*, Dublin, Ireland, June 28–30, pp. 521–526.

Werker, J. F., & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. *Language Learning and Development*, 1, 197–234.