Comparison of Syllabification Algorithms and Training Strategies for Robust Word Count Estimation across Different Languages and Recording Conditions

Okko Räsänen[1], *Shreyas Seshadri*[1], *Marisa Casillas*[2]

[1]Department of Signal Processing and Acoustics, Aalto University, Finland
[2]Max Planck Institute for Psycholinguistics, Netherlands

okko.rasanen@aalto.fi, shreyas.seshadri@aalto.fi, marisa.casillas@mpi.nl

Abstract

Word count estimation (WCE) from audio recordings has a number of applications, including quantifying the amount of speech that language-learning infants hear in their natural environments, as captured by daylong recordings made with devices worn by infants. To be applicable in a wide range of scenarios and also low-resource domains, WCE tools should be extremely robust against varying signal conditions and require minimal access to labeled training data in the target domain. For this purpose, earlier work has used automatic syllabification of speech, followed by a least-squares-mapping of syllables to word counts. This paper compares a number of previously proposed syllabifiers in the WCE task, including a supervised bi-directional long short-term memory (BLSTM) network that is trained on a language for which high quality syllable annotations are available (a "high resource language"), and reports how the alternative methods compare on different languages and signal conditions. We also explore additive noise and varying-channel data augmentation strategies for BLSTM training, and show how they improve performance in both matching and mismatching languages. Intriguingly, we also find that even though the BLSTM works on languages beyond its training data, the unsupervised algorithms can still outperform it in challenging signal conditions on novel languages.

Index Terms: language acquisition, syllabification, word count estimation, daylong recordings, noise robustness

1. Introduction

Automatic word count estimation (WCE) is the task of estimating the total number of words spoken in an audio recording. For instance, it can be used to investigate social dynamics or speaker characteristics from large-scale audio recordings. Typical use cases include signal conditions, low-resource domains, or computationally constrained environments where standard automatic speech recognition (ASR) may not be applicable or may not be the best performing solution for the task (see also [1] for an overview).

One central use case for WCE is the study of infant language learning, where daylong recordings from infants' natural environments are collected in order to study the quality and quantity of language input that the children hear (and later produce) in their normal daily lives (e.g., [2,3]). For instance, our ongoing *Analyzing Child Language Experiences Around the World* (ACLEW) project [4] aims to quantify how much speech children hear in a variety of cultural and socioeconomic environments. The goal is to understand how aspects of early language experience map to later developmental outcomes. A comparative project of this sort calls for efficient and unbiased WCE methods that can be used across a number of language and recording environments.

In order to achieve comparable results for both high- and low-resource languages, and to function in signal conditions encountered in daylong recordings such as those collected with the wearable LENA® recording system [5], the WCE tools should be extremely robust against different types of noise. They should also be sensitive to near- and far-field speech in different indoor- and outdoor-environments, and require only minimal access to labeled training data in the target domain. In addition, the massive scale of the daylong recording datasets—often accumulating to hundreds or even thousands of hours of audio—imposes limitations to what can be computed in a reasonable time on non-dedicated computing environments (e.g., linguists or psychologists using the tools in Speech Recognition Virtual Kitchen described in [6], through which our ultimate WCE solution will be distributed).

In earlier work on WCE and the closely-related task of speech rate estimation (SRE), the predominant approach has been to first estimate the number of syllables [1,7–9] or ASR-based vowels and consonants [10] in a stretch of speech. A linear mapping from these units to the corresponding utterance-level word counts or speaking rate is then performed (but see also [11]). Since syllable nuclei correspond to local sonority peaks in the speech signal, the syllabification process can be carried out in an unsupervised manner (e.g., [12–16]). In addition, it is, in principle, applicable to any language and is robust against signal degradations due to the energetic nature of syllabic nuclei (see also [16] and references therein).

Notably, the studies on WCE and SRE have all used either English [1, 7–11] or Dutch [14] speech only, and use matching corpora to develop, tune, and compare the systems. In addition, the existing WCE systems largely rely on unsupervised syllabification algorithms. Recently, a study in [17] reported promising results on the use of a supervised bi-directional long short-term memory (BLSTM) network for syllable-based SRE. However, the method was only tested on English data. Whether the approach would also be applicable to WCE in low-resource domains, where syllabic annotations may not be available for model training, is unclear.

In this paper, we go beyond the existing WCE studies by 1) comparing three previously successful syllabification methods on multiple corpora, languages, and varying signal conditions, and 2) investigating the applicability of a supervised BLSTM syllabifier to mismatching languages, exploring training data augmentation strategies to improve its performance. The overall aim is to understand what type of syllabification approach would be the most applicable one to be used as an off-the-shelf solution for WCE.

2. Data

Five speech corpora were used in the experiments to compare syllable estimation methods across languages and signal conditions: **1) FinDialogue Corpus of Spontaneous Finnish Speech** [18], having studio quality dialogues from 4 talkers (2 male; 64 minutes of speech in total), **2) Phonetic Corpus of Estonian Spontaneous Speech** ("EstPhon"; [19]), consisting of a studio dialogue section of 21 talkers (310 min) used for BLSTM training, and a fieldwork recording section of 10 talkers (100 minutes) collected with headsets and used for method comparisons, **3) Switchboard** ("SWB") corpus of spontaneous American English telephone conversations [20], using a syllable-annotated ICSI transcription project subset (>10 talkers, 159 minutes), and **4) Brent&Siskind corpus** ("Brent") of American English infant-directed speech [21], using the so-called "Large Brent" subset force-alignment annotated in [22] (4 talkers, 103 min). All these syllable-annotated subsets of the full corpora were used since we were also interested in syllabification performance (not reported here; see also [16] for details on the data).

In addition, we used extracts from **5) daylong recordings made with children from a rural Tseltal Mayan community**, collected by the third author using lightweight stereo digital voice recorders (either an Olympus® WS-832 or WS-853) that was worn on the target child's chest in an elastic vest [23]. The current subset consists of recordings from 10 children between 2- and 36-months old who live in households with 3 to 11 other people. From the original 10–11 hour daylong recordings of each child, nine randomly sampled five-minute chunks were manually annotated for utterance boundaries, orthographic transcription, and talker identity information. In the present experiments, we included all utterances from all male and female adult talkers for which the transcripts were fully unambiguous in terms of the number of words spoken, corresponding to 6458 utterances (191 min of speech out of 450 min total audio). Note that the dataset still contains overlapping speakers, background noise from the ambient environments, and mainly far-field speech (often in reverberant environments), making these data extremely challenging in comparison to standard corpora.

Overall, the data represents three rhythmic families (stress-timed English, syllable-timed Finnish and Tseltal, mixed-timed Estonian), adult- and infant-directed speech, and signal qualities ranging from studio quality to highly noisy daylong recordings from wearable microphones.

3. Methods

The basic WCE pipeline in the present study follows that of [1], and is shown at the top of Fig. 1. A complete WCE system would typically be expected to work on utterances extracted using a speech activity detection (SAD) algorithm. However, here we examine the performance of different WCE methods under ideal segmentation by using true utterance boundaries from manual annotation. This lets us to compare performance without added SAD error and, additionally, allows us to benchmark on corpora with readily segmented utterances.

In the present WCE, the speech signal $\mathbf{x}_u$ for utterance u is first transformed into a syllabic ("sonority") envelope $\mathbf{y}_u$ using one of three alternative methods: 1) *thetaSeg* that was originally developed to study infant syllable segmentation using perceptually motivated entrainment to sonority fluctuations in speech [16], 2) A method by Wang & Narayanan [8] ("*WN*"), originally developed for SRE and

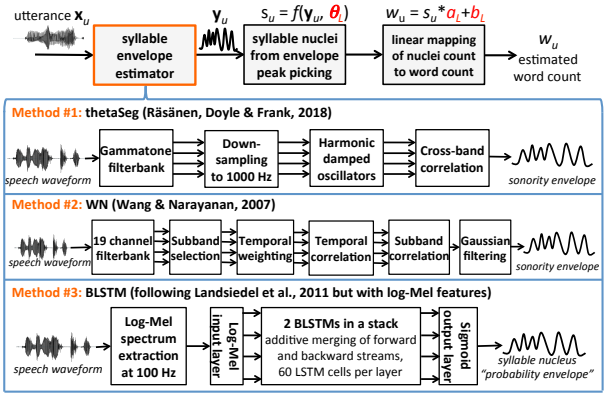

Figure 1: *Block schematics of the WCE system (top) and the three primary syllable envelope estimators compared in the experiments. Threshold θ_L and LMSE parameters estimated for each target language L are shown with red font.*

optimized for SWB, and 3) a BLSTM network following the approach in [17]. However, we replaced the original modulation spectrum and PLP feature set with 24-channel log-Mel features, having four hidden LSTM layers instead of two (2 forward, 2 backward), and doubling the number of LSTM cells to 60. These changes were done to provide more capacity for representation learning in the hidden layers, as this was found to improve performance in our initial experiments. Basic processing steps in each of the three methods can be found in Fig. 1, and implementation details are available in the respective papers. WN was used with the default parameters optimized for Switchboard in [8], while thetaSeg was used with the optimal parameters for multi-language syllable segmentation described in [16]. All envelopes were linearly scaled to have maximum value of 1 for each utterance to ensure consistency across signal conditions and sound levels, assuming that the maximum peak in the envelope corresponds to the most sonorous speech sound in the signal.

Syllable envelope estimation is followed by syllable count estimation using a peak-picking algorithm that has a language-specific threshold $\theta_L \in [0, 1]$ for the required difference between the peak and the previous minimum, thereby controlling the sensitivity of the process. The resulting syllable counts s_u are then mapped to word counts using a linear model whose slope a_L and intercept b_L are also optimized for the target language L. The three parameters (θ_L, a_L, b_L) are jointly optimized for L by testing the full range of θ_L values and finding the best least-squares linear mapping (LMSE) for each threshold based on RMSE word count error across all utterances on orthographically transcribed training data for L. The best parameter triplet is then chosen for that language and used for estimating word counts on novel test data.

Note that all three syllabifiers are used without their additional mechanisms for pruning erroneous nuclei, as we focus on comparing envelope estimation techniques with a common envelope decoding stage, and since some of the techniques (e.g., F0-based pruning in WN) are difficult to tune properly for generic use across corpora and SNRs. Therefore our results will only reflect the envelope representations of the compared methods, not the full pipelines reported in the original papers.

Also note that, for infant daylong recordings, the typical time-scales of interest can be the number of words heard during an interaction scenario, across some hours, or even over a full day. Therefore the key for good performance in the

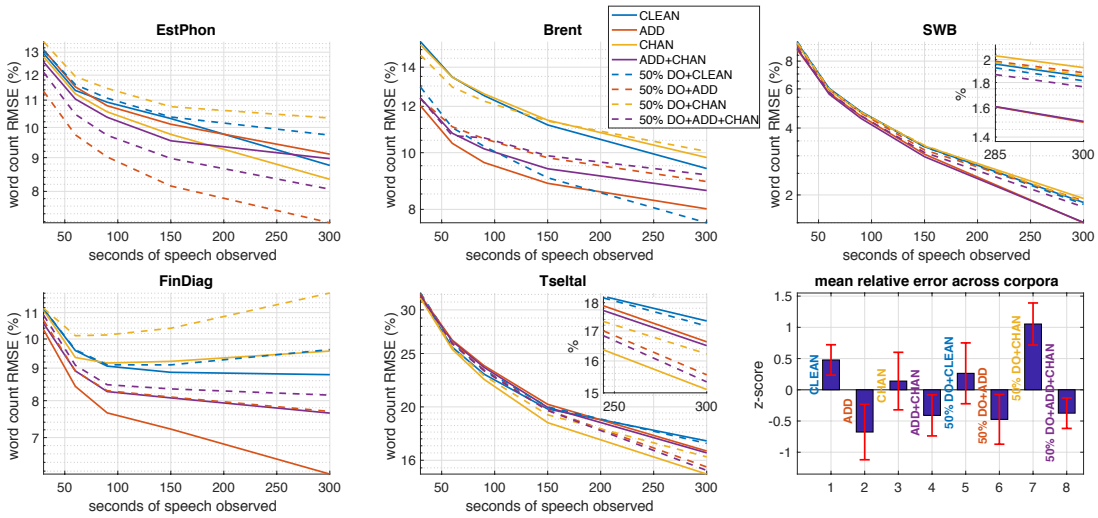

Figure 2: *Performance of alternative BLSTM training strategies as a function of the amount of speech observed. Bottom right panel shows the overall relative performance of each variant across the five corpora (at 300 s), where word RMSEs (%) have been z-score normalized across the BLSTM variants before averaging (smaller is better). Horizontal bars denote SEs over the corpora.*

present type of WCE pipeline is to have a *robust* and *systematic* syllable envelope estimator across speakers and recording environments. The estimator can under-segment or over- segment as long as the behavior is consistent over the time-scales of interest. Any constant errors can be overcome during the linear mapping from the syllable counts to word counts, assuming that the mapping itself is also unbiased. As the length of the estimation window increases, any unbiased estimator will eventually converge to zero error.

3.1. BLSTM training and data augmentation

The BLSTM training was carried out similarly to [17] but now using the studio section of the EstPhon corpus with clean wide-band speech from multiple talkers and high-quality manual labels for syllables and phones. Syllable annotation of the training data was used to create target output signals at the same frame rate (100 Hz) as the input log-Mel features. A Gaussian-shaped pulse with maximum amplitude of 1 was centered on each syllable nucleus with its standard deviation corresponding to one third of phone duration at the nucleus position, other signal values being 0. The BLSTM was then trained with RMSprop optimizer [24] to minimize the error in mapping log-Mel features into the syllabic targets using binary cross-entropy loss, early stopping, and minibatch size of 100.

To improve generalization beyond clean Estonian speech ("CLEAN"), we tested additive noise-augmented ("ADD"), channel-augmented ("CHAN"), and ADD&CHAN-augmented training (combined in the given order). For ADD, we sampled random segments from daylong recordings taken from the ACLEW starter set [25] that could contain any sounds from the at-home environments of infants across the world. The noise was then added to the original utterances with SNR randomly sampled from [-10, 20] dB, three copies of the signal corrupted with different random noises. For CHAN, we created three copies of each utterance, each filtered with a 10-point FIR filter (at 16 kHz) with filter coefficients randomly sampled from a normal distribution ~$N(0, 1)$, as this produced perceptually sensible variation to the channel properties.

We also tested training without dropout and with 50% dropout in the hidden layers following the merging of forward and backward streams. The results reported for each BLSTM variant correspond to the average performance across three identically specified models trained with random parameter initialization for each, thereby reflecting the expected performance in new domains without further model selection.

4. Experiments

4.1. Experimental setup

N-fold evaluation procedure was used for all corpora, always using data from N-1/N of the talkers (or recorded infants for Tseltal and Brent) to estimate θ_L, a_L, and b_L for the fold, and testing generalization to the last 1/N of the subjects. The number of folds was set to 10 for Tseltal and to 4 for Brent and FinDiag to match the number of unique subjects. Six folds were used for the EstPhon fieldwork section and Switchboard, which had many more subjects (section 2). Following [10], WCE accuracy was measured for 30-, 60-, 90-, 150-, and 300-second chunks of speech. The corresponding estimated and true word counts were obtained by concatenating outputs for temporally subsequent utterances until the desired signal duration was reached. We report the average word count RMSEs across the folds for these time-scales.

As a baseline, we also computed duration-based results simply by replacing the nuclei counts s_u with the corresponding utterance durations in the pipeline (see Fig. 1).

4.2. Results

Figure 2 shows the results of the BLSTM variants for each five corpora as a function of the amount of speech observed. In addition, the average relative performance of each variant across all corpora is shown in the bottom right panel. The numbers in the final panel correspond to averages of corpus-specific z-score-normalized RMSEs (%) at 300 s of speech in order to normalize the scale of errors across corpora.

As can be observed, the best training strategy depends on the corpus in question. On average, combined additive noise and channel augmentation outperforms clean training with and without dropout, but using only additive noise seems to work well in most cases. Interestingly, the channel augmented training works the best for the most difficult scenario, the Tseltal corpus. It also helps on Estonian field recordings

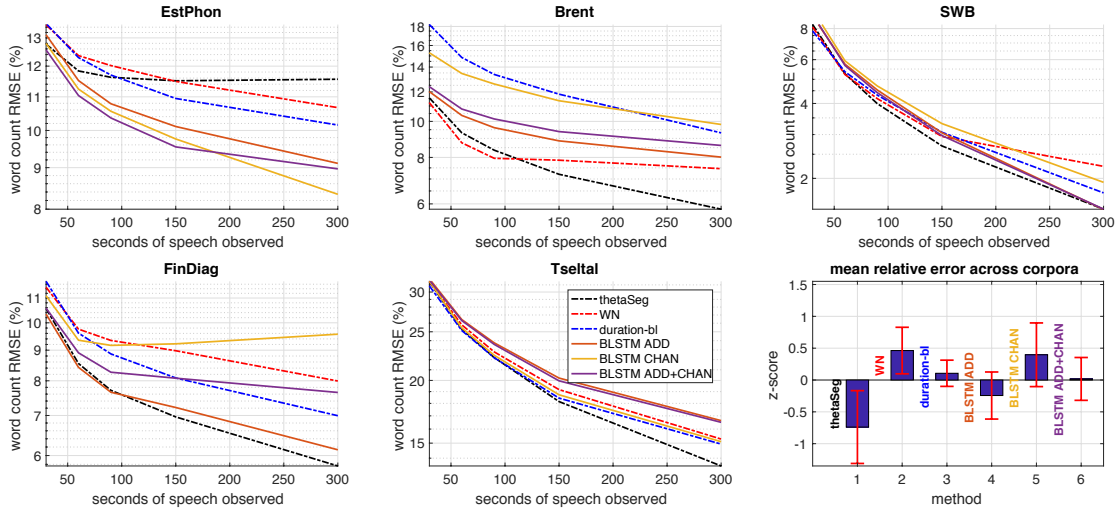

Figure 3: *Performance of the baseline systems and three main BLSTM variants in different corpora as a function of the amount of speech observed. Bottom right panel shows the overall relative performance of each variant across the five corpora (at 300 s).*

compared to clean training, whereas its benefits for the other corpora are more limited. As for dropout, the results are surprisingly mixed, considering that each BLSTM variant was re-trained three times and results are averaged across these runs. In case of clean training, dropout always helps except for FinDiag. In case of noise-augmented training, dropout has substantial benefits on the matching Estonian language with new speakers and recording environments. However, for other corpora, dropout does not help when augmentation is used.

Figure 3 shows the corresponding results for the two unsupervised methods together with the three main BLSTM variants using data augmentation (no dropout) and the duration baseline. As can be observed, all supervised BLSTMs clearly outperform the unsupervised approaches on the EstPhon corpus that matches the training language of the BLSTMs. In addition, BLSTM performance on Switchboard is among the best, and would likely win at longer time-scales of analysis. However, the comparison also reveals that the unsupervised syllable envelope estimators also perform well. The most consistent performer at longer estimation windows on all but Estonian data is the thetaSeg algorithm. On EstPhon, thetaSeg seems to have problems in finding an unbiased mapping from syllable nuclei to word counts, visible as the lack of improvement as a function of more speech data. WN also works well on Switchboard, Brent, and Tseltal at shorter time-scales, but, for currently unknown reasons, does not converge towards zero-error as quickly as the thetaSeg as more speech is observed, suggesting that the errors made by WN are more correlated with each other than in thetaSeg.

Another interesting finding is that the duration-based baseline performs relatively well, especially on Tseltal, but also to some degree on Switchboard. While the former can be explained by the extremely challenging signal conditions where accurate syllable detection is prone to errors due to overlapping sounds and generally poor SNR, the latter could be due to the relatively regular syllable and word rates in telephone conversations compared to, e.g., the Brent data (everyday communication between infants and caregivers), which has very low accuracy for the duration baseline.

5. Conclusions

In a first effort to compare existing WCE systems across a diverse range of language and recording contexts, the present results suggest that a supervised BLSTM syllabifier with training data augmentation has the potential to be used for WCE across languages and mismatching signal conditions. We also demonstrated that a simple random-FIR approach to varying-channel data augmentation can improve generalization performance from clean training, suggesting that such an approach could also be useful for other speech technology applications. That said, the results also show that despite the aggressive overfitting prevention through noise augmentation and dropout training, unsupervised approaches to syllabification still outperform the BLSTM in languages for which syllable-level training data are not available.

However, it is also clear that the BLSTM approach is highly recommendable as long as data from a matching language are available for training of the system, supporting the findings of Landsiedel et al. [17]. BLSTM is also potentially more flexible than the compared unsupervised systems in terms of adapting the model to new domains and languages, as long as some kind of syllabic annotations or their proxies can be derived for the target domain. Thus, future work should investigate different adaptation or even end-to-end strategies for the BLSTM, given access to orthographic information that is needed anyway to tune the language-specific mapping parameters. Multi-language training of the BLSTM should also be investigated, given suitable corpora with high-quality syllable annotations in different languages.

One limitation of the present study is that all experiments, although necessary for the present cross-corpus comparisons, assume ideal segmentation of speech into utterances. In actual WCE scenarios with daylong recordings, SAD-based utterance segmentation is likely to contain errors. Tolerance of the present algorithms against such errors is currently unknown, and should be investigated on the daylong data in the future.

6. Acknowledgements

This research was funded as a part of *Analyzing Child Language Experiences around the World* (ACLEW) collaboration grant funded by the Trans-Atlantic Platform for Social Sciences and Humanities "Digging into Data" challenge, an Academy of Finland grant (312105) to OR, and an NWO Veni Innovational Research grant (275-89-033) to MC. The authors would like to thank Elika Bergelson for the useful feedback on the manuscript.

7. References

[1] A. Ziaei, A. Sangwan, and J. Hansen, "Effective word count estimation for long duration daily naturalistic audio recordings," *Speech Communication*, vol. 84, pp. 15–23, 2016.

[2] M. Soderstrom, and K. Wittebolle, "When do caregivers talk? The influences of activity and time of day on caregiver speech and child vocalizations in two childcare environments," *PLoS One*, vol. 8, no. 11:e80646, 2013.

[3] A. Weisleder, and A. Fernald, "Talking to children matters: Early language experience strengthens processing and builds vocabulary," *Psychological Science*, vol. 24, no. 11, pp. 2143–2152, 2013.

[4] M., Soderstrom, E. Bergelson, C. Rosemberg, E. Dupoux, O. Räsänen, B. Schuller, A. Cristia, F. Metze, F. Rudzicz, and M. Casillas, "*Analyzing the Child Language Experiences around the World Project*". https://sites.google.com/view/aclewdid/home

[5] J. Gilkerson, and J. Richards, "The LENA Natural Language Study," *LENA Foundation Technical Reports* (September 2008), pp. 1–26, 2009.

[6] F. Metze, E. Fosler-Lussier, and R. Bates, "The Speech Recognition Virtual Kitchen," *Proc. Interspeech-2013*, Lyon, France, August 25–29, 2013, pp. 1858–1860. https://github.org/srvk.

[7] N. Morgan, and E. Fosler-Lussier, "Combining multiple estimators of speaking rate," *Proc. ICASSP-1998*, Seattle, Washington, May 12–15, 1998, vol. 2, pp. 729–732.

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

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

[10] D. Xu, U. Yapanel, S. Gray, J. Gilkerson, J. Richards, and J. Hansen, "Signal processing for young child speech language development," *Proc. WOCCI-2008*, Crete, Greece, October 23, 2008.

[11] Y. Jiao, V. Berisha, M. Tu, and J. Liss, "Convex Weighting Criteria for Speaking Rate Estimation," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1421–1430, 2015.

[12] 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.

[13] P. Mermelstein, "Automatic segmentation of speech into syllabic units," *Journal of the Acoustical Society of America*, vol. 58, pp. 880–883, 1975.

[14] N. 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.

[15] N. Obin, F. Lamare, and A. Roebel, "Syll-O-Matic: An adaptive time-frequency representation for the automatic segmentation of speech into syllables," *Proc. ICASSP-2013*, Vancouver, Canada, May 26–31, 2013, pp. 6699–6703.

[16] O. Räsänen, G. Doyle, and M. C. Frank, "Pre-linguistic segmentation of speech into syllable-like units," *Cognition*, vol. 171, pp. 130–150, 2018.

[17] C. Landsiedel, J. Edlund, F. Eyben, D. Neiberg, and B. Schuller, "Syllabification of conversational speech using bidirectional long-short-term memory neural networks," *Proc. ICASSP-2011*, Prague, Czech Republic, May 22–27, 2011, pp. 5256–5259.

[18] M. Lennes, "Segmental features in spontaneous and read-aloud Finnish," In: V. de Silva & R. Ullakonoja (Eds.): *Phonetics of Russian and Finnish*, pp. 145–166, Frankfurt am Main: Peter Lang, 2009.

[19] P. Lippus, T. Tuisk, N. Salveste, and P. Teras, "Phonetic corpus of Estonian spontaneous speech," Institute of Estonian and General Linguistics, University of Tartu, 2013. DOI: https://doi.org/10.15155/TY.000D.

[20] J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: telephone speech corpus for research and development," *Proc. ICASSP-1992*, San Francisco, CA, March 23–26, 1992, pp. 517–520.

[21] M. Brent, and J. Siskind, "The role of exposure to isolated words in early vocabulary development," *Cognition*, vol. 81, 31–44, 2001.

[22] C. A. Rytting, C. Brew, and E. Fosler-Lussier, "Segmenting words from natural speech: subsegmental variation in segmental cues," *Journal of Child Language*, vol. 27, no. 03, pp. 513–543, 2010.

[23] M. Casillas, P. Brown, and S. C. Levinson, "Casillas HomeBank Corpus", 2017. https://homebank.talkbank.org/access/Secure/Casillas.html

[24] G. Hinton, N. Srivastava, and K. Swersky, "Neural Networks for Machine Learning: Lecture 6a, Overview of mini-batch gradient descent," Coursera online lecture slides, 2012.

[25] E. Bergelson, A. Warlaumont, A. Cristia, M. Casillas, C. Rosemberg, M. Soderstrom, F. Metze, E. Dupoux, O. Räsänen, C. Rowland, and S. Durrant. Starter-ACLEW. *Databrary*. Retrieved February 28, 2018 from http://doi.org/10.17910/B7.390, 2017.