Räsänen, Okko; Seshadri, Shreyas; Karadayi, Julien; Riebling, Eric; Bunce, John; Cristia, Alejandrina; Metze, Florian; Casillas, Marisa; Rosemberg, Celia; Bergelson, Elika; Soderstrom, Melanie

Automatic word count estimation from daylong child-centered recordings in various language environments using language-independent syllabification of speech

Published in:
Speech Communication

DOI:
10.1016/j.specom.2019.08.005

Published: 01/10/2019

Document Version
Publisher's PDF, also known as Version of record

Speech Communication 113 (2019) 63–80

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Automatic word count estimation from daylong child-centered recordings in various language environments using language-independent syllabification of speech

Okko Räsänen[a,b,*], Shreyas Seshadri[b], Julien Karadayi[c], Eric Riebling[d], John Bunce[e], Alejandrina Cristia[c], Florian Metze[d], Marisa Casillas[f], Celia Rosemberg[g], Elika Bergelson[h], Melanie Soderstrom[e]

[a] Unit of Computing Sciences, Tampere University, P.O. Box 553, FI-33101 Tampere, Finland
[b] Department of Signal Processing and Acoustics, Aalto University, Finland
[c] Laboratoire de Sciences Cognitives et Psycholinguistique, Dept d'Etudes Cognitives, ENS, PSL University, EHESS, CNRS, France
[d] Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA
[e] Department of Psychology, University of Manitoba, Canada
[f] Max Planck Institute for Psycholinguistics, The Netherlands
[g] Centro Interdisciplinario de Investigaciones en Psicología Matemática y Experimental, CONICET, Argentina
[h] Department of Psychology and Neuroscience, Duke University, NC, USA

ARTICLE INFO

Keywords:
Language acquisition
Word count estimation
Automatic syllabification
Daylong recordings
Noise robustness

ABSTRACT

Automatic word count estimation (WCE) from audio recordings can be used to quantify the amount of verbal communication in a recording environment. One key application of WCE is to measure language input heard by infants and toddlers in their natural environments, as captured by daylong recordings from microphones worn by the infants. Although WCE is nearly trivial for high-quality signals in high-resource languages, daylong recordings are substantially more challenging due to the unconstrained acoustic environments and the presence of near- and far-field speech. Moreover, many use cases of interest involve languages for which reliable ASR systems or even well-defined lexicons are not available. A good WCE system should also perform similarly for low- and high-resource languages in order to enable unbiased comparisons across different cultures and environments. Unfortunately, the current state-of-the-art solution, the LENA system, is based on proprietary software and has only been optimized for American English, limiting its applicability. In this paper, we build on existing work on WCE and present the steps we have taken towards a freely available system for WCE that can be adapted to different languages or dialects with a limited amount of orthographically transcribed speech data. Our system is based on language-independent syllabification of speech, followed by a language-dependent mapping from syllable counts (and a number of other acoustic features) to the corresponding word count estimates. We evaluate our system on samples from daylong infant recordings from six different corpora consisting of several languages and socioeconomic environments, all manually annotated with the same protocol to allow direct comparison. We compare a number of alternative techniques for the two key components in our system: speech activity detection and automatic syllabification of speech. As a result, we show that our system can reach relatively consistent WCE accuracy across multiple corpora and languages (with some limitations). In addition, the system outperforms LENA on three of the four corpora consisting of different varieties of English. We also demonstrate how an automatic neural network-based syllabifier, when trained on multiple languages, generalizes well to novel languages beyond the training data, outperforming two previously proposed unsupervised syllabifiers as a feature extractor for WCE.

1. Introduction

Automatic word count estimation (WCE) from audio recordings can be used to investigate vocal activity and social interaction as a function of recording time and location, such as in personal life logs derived from wearable sensors (Ziaei et al., 2015, 2016). WCE is also a highly useful tool in the scientific study of child language acquisition because it can help answer questions such as how much speech children hear in their

* Corresponding author at: Unit of Computing Sciences, Tampere University, P.O. Box 553, FI-33101 Tampere, Finland.
E-mail address: okko.rasanen@tuni.fi (O. Räsänen).

https://doi.org/10.1016/j.specom.2019.08.005
Received 5 December 2018; Received in revised form 5 August 2019; Accepted 13 August 2019
Available online 14 August 2019

daily lives in different contexts (e.g., Bergelson et al., 2018a), and how the language input maps to later developmental outcomes in the same children (Weisleder and Fernald, 2013; Ramírez-Esparza et al., 2014). In the present work, we focus on the latter application.

It is already known that there are substantial differences in language exposure between families, socioeconomic environments, and cultures, with potential impact on later language development outcomes (Hart and Risley, 1995; Huttenlocher et al., 2010; Rowe, 2012; Weisleder and Fernald, 2013; see also Hoff, 2006, for a review). Such differences may relate to the quantity and kind of speech children hear, but also to questions such as how often the infant is addressed directly, and how often they overhear adult conversations (e.g., Lieven, 1994; Shneidman and Goldin-Meadow, 2012; Cristia et al., 2017). However, many of these conclusions have been drawn from short observations of child-caregiver interaction recorded in a lab or at the child's home, providing only a limited view into the daily variation children encounter in their linguistic input (Tamis-LeMonda et al., 2017; Bergelson et al., 2018b). Furthermore, the vast majority of this research has been carried out in the context of limited set of languages and cultural environments, largely focusing on so-called WEIRD communities (Western, Educated, Industrialized, Rich, Democratic; Henrich et al., 2010) which limits the generalizability of the findings. To better study the input and its effects on development, and in response to changing technological availability, language development researchers have increasingly been recording children as they go about their daily lives with wearable microphones, allowing quantification of language input from data corresponding to the natural learning environments of the children. However, since it is not realistic to manually annotate hundreds or even thousands of hours of audio data from such daylong recordings, automated speech processing solutions are needed. This is where automatic WCE systems can come to the rescue, as they can provide an invaluable automated tool for measuring the number of words children have heard in a period of time.

The existing state-of-the-art solution for the daylong recording and analysis task is the LENA^{TM} system (Xu et al., 2008; Gilkerson and Richards, 2009) developed by the LENA Research Foundation (http://www.lena.org). The LENA setup includes a compact recorder that can be placed inside the pocket of a vest worn by the child, and software that analyzes various aspects of the child's daily language experience from the audio, including measures such as conversational turns, adult word counts, and counts of child vocalizations. Despite its tremendous value for the language research community, LENA as a software solution is not without problems. First, the software is proprietary and expensive. Second, only audio captured with the LENA recorder can be analyzed with the software, i.e. other audio files cannot be run through the same software. In addition, the included algorithms for speech processing are potentially outdated due to aging of the system, the basic building blocks having been introduced nearly 10 years ago (e.g., Xu et al., 2008). Since the algorithms are not open-source, it is also not possible to improve the software. Finally, LENA speech processing algorithms, including the WCE module, have been optimized for American English. While the system can be used with recordings in any language, its accuracy is not necessarily consistent across different populations, complicating any attempt at cross-linguistic comparison.

Given this background, there is an increasing demand from the research community to develop an alternative to LENA that would be open source, free of charge, compatible with audio data obtained using a variety of recorders, and robustly applicable to a variety of languages and language environments. In order to address this challenge, our ongoing collaborative project called *Analyzing Child Language Experiences Around the World* (ACLEW; https://sites.google.com/view/aclewdid/home) aims to develop an open-source software package that would address the mentioned shortcomings of LENA (see Le Franc et al., 2018, for initial work). The developed tools will be distributed as a Linux virtual machine that can be operated on a variety of computing platforms without special technical expertise in installing or operating speech processing algorithms (see Metze et al., 2013; Plummer et al., 2014). The system also aims to be scalable to large data sets with modest computational resources, as the aim is to make the tools usable by a broad population of researchers using a variety of computing environments. WCE is one among several tools under development that we hope to integrate into the software package.

In the present paper, we describe our recent developments for the WCE component of the daylong analysis toolkit. After describing the WCE problem in more detail (the next subsection), we present our basic WCE pipeline. In a nutshell, our solution is based on language-independent syllabification of speech, followed by a language-dependent mapping from syllable counts (and a number of other acoustic features) to the corresponding word count estimates. Our work extends the earlier WCE system by Ziaei et al. (2016) and also earlier syllable-based speech rate estimators such as those by Wang and Narayanan (2007) and Morgan and Fosler-Lussier (1998). However, we go beyond the existing studies by (1) investigating applicability of syllable-based WCE to daylong child-centered recordings in several languages and in participant samples with varied socioeconomic status, and (2) comparing the impact of several speech activity detectors (SADs) and syllabifiers on the WCE performance. In addition, we (3) explore cross-language generalization of a language-independent supervised syllabification algorithm, thereby potentially replacing the unsupervised syllabification algorithms (Ziaei et al., 2016) or acoustic phone models (Xu et al., 2008) used in the earlier WCE systems. The ultimate aim of this study is to identify the best performing WCE system configuration that generalizes well to new languages and domains, and to see how it compares against LENA performance.

1.1. The WCE problem

The key idea of a WCE system is to infer the number of spoken words in a given audio signal segment. Ideally the word count estimates would already be accurate at the level of individual utterances. However, due to the extremely challenging signal conditions encountered in typical daylong recordings, this turns out to be a difficult problem in practice. Because the recording device (e.g., the LENA recorder) is worn by the child and records continuously, the microphone picks up not only speech of the child and caregivers, but also any other audible sounds in the environment. These sounds can include varying ambient noises, overlapping speech, and non-linguistic vocalizations. Moreover, each sound source (including speech of interest) has different channel characteristics due to the varying geometries of the spaces and source positions. In addition, signal artefacts from clothing scratching against the microphone during child movement are also common. A large proportion of the collected data is also mono (e.g., all LENA output), removing any useful directional information that could help in source separation. Finally, researchers are increasingly collecting daylong recordings with a variety of non-LENA recorders, which means the technical characteristics of the devices can also differ from one dataset to the next. This means that the overall properties of the audio data are largely uncontrolled, calling for robust signal processing methods.

Another central challenge comes from the cross-domain applicability of the WCE system: performance of the system should ideally be similar in high-resource languages such as English (across all its dialects and social environments of the talkers) and in low-resource languages, such as Tseltal, a Mayan language included in our experiments. In conjunction with the problematic signal conditions, this limits the applicability of standard ASR systems for WCE in cross-linguistic developmental research. Balancing the performance of language-specific ASR systems for the different language environments is not trivial, especially considering the challenges involved in obtaining sufficiently representative lexicons, pronunciation dictionaries, and language models for low-resource languages.

Fortunately, the use of WCE for developmental research may not require systems to identify individual words from the speech stream,

Table 1
A list of LENA word count estimation accuracies reported in the literature, as measured between LENA output and manually annotated word counts. *N* denotes the total number of samples used in performance calculation and "segment duration" refers to the duration of audio in each sample (there can be one or more samples per subject). All reported mean (ERR_{mean}) and median (ERR_{median}) absolute relative errors across subjects have been derived by the present authors from the word count data reported in the publications, or from data obtained from the original authors of the studies through personal communication. See also Section 4.1 for details on use of Eqs. (1) and (2).

Authors	Language	*r*	ERR_{mean}	ERR_{median}	*N*	Segment duration	Other notes
Xu et al. (2009)	American English	0.92	N/A	N/A	70	1 h	Data sampling not specified, but most likely the same 1 h segments with high speech activity as in Xu et al. (2008).
Soderstrom and Wittebolle (2013)	Canadian English	0.76	34.1%	27.7%	10	100 min	
Canault et al. (2016)	French	0.64	177.0%	36.5%	324	10 min	Hand-picked segments with high vocal interaction
Canault et al. (2016)	French	0.37	31.2%	27.5%	18	3 h	Same as above with data pooled across 18 subsequent sessions across several days.
Weisleder and Fernald (2013)	Mexican Spanish	0.80	45.2%	50.2%	10	1 h	
Schwarz et al. (2017)	Swedish	0.67	78.4%	59.5%	48	5 min	Only 4 subjects.
Schwarz et al. (2017)	Swedish	0.86	42.8%	37.0%	4	1 h	Only 4 subjects.
Elo (2016)	Finnish	0.99	75.2%	55.3%	21	1 h	Only 2 subjects.
Gilkerson et al. (2015)	Shanghai and Mandarin Chinese	0.73	N/A	N/A	22	15 min	
Busch et al. (2018)	Dutch	0.88	496.6%	42.9%	65	5 min	Derived from 65 × 5 min samples provided by Busch instead of the 48 samples in the original study.
Busch et al. (2018)	Dutch	0.92	32.7%	34.2%	6	54 min	Above data, but pooled across all 5 min segments per subject (4–16 segments, 54 min average total duration).

enabling alternative technological solutions. A typical use case may be concerned with questions such as "*How many words did this child hear per day*?" (e.g., Weisleder and Fernald, 2013) or "*How many words does the child hear at day care* versus *at home?*" (e.g., Soderstrom and Wittebolle, 2013), and where such aggregate word counts are then related to other variables of interest. This means that the relevant time-scales are often measured in terms of several minutes, if not hours or days, instead of individual utterances or words. This enables the use of statistical approaches to WCE where estimates of aggregate word counts can be derived from features or representations of the signal that, *on average*, depend on the number of words in the data. For instance, the LENA WCE module first detects the total number of vowels and consonants in the signal using an acoustic phone model, and combines these with measured speech duration with and without silences (Xu et al., 2008). These features (and their square roots) are then mapped to the expected corresponding word count using a least-squares linear mapping that has also been optimized on American English. Another WCE system recently proposed by Ziaei et al. (2016) takes a similar approach. However, instead of phone counts their system uses syllable counts from an unsupervised syllabifier by Wang and Narayanan (2007) as the primary feature.

In both the phone and syllable-based WCE systems above, the key assumption is that speakers of the given language share a lexicon that is stationary in terms of *average* phonemic or syllabic length of words at the time-scales of interest. Even though a system might not get the estimated word count right for individual utterances (since it does not identify individual word forms as such), the estimation error will converge to zero over time as long as the estimator is unbiased, i.e., as long as the system does not systematically under- or overestimate the word counts at the utterance-level. In this context, short-term accuracy of the estimator will simply determine the rate at which the estimation error decreases when more speech is observed. Given unbiased estimators with a sufficient accuracy, a WCE system may therefore provide useful word count estimates at the time-scales of interest, even if it does not know the lexical or phonological properties of the language in detail. This is also a property that we utilize in our present system, as will be described in Section 2.

1.2. State-of-the-art, open issues, and the present contributions

So far, there are essentially two systems for WCE that have been proposed in the earlier literature, LENA and the system by Ziaei et al. (2016), both already mentioned above. While LENA is specifically designed for analyzing child-centered daylong recordings (including a WCE module for measuring speech *heard* by the infant), the system by Ziaei et al. was designed to only count the words of the person wearing the microphone. Their best performing system variant uses TO-Combo-SAD (Sadjadi and Hansen, 2013; Ziaei et al., 2014) for speech detection, spectral subtraction for speech enhancement, and an automatic syllabifier from Wang and Narayanan (2007) for syllable count estimation before mapping the counts to word counts. Ziaei et al. evaluated their system on Prof-Life-Log database consisting of 13 recordings from one adult participant wearing the LENA microphone during typical working days, with the data manually transcribed for word counts (Ziaei et al., 2016; see also Ziaei et al., 2013, 2014). According to our knowledge, applicability of their system to child-centered daylong data has not been tested to date. However, our experiments will partially address this issue by having one of our WCE system configurations being highly similar to theirs (i.e., using TO-Combo-SAD, spectral subtraction, the same syllabifier, and a linear model between signal features and words). Also, our system generally builds on that work, but as we show in the experiments, we also introduce more robust techniques for automatic syllabification of speech in daylong recording conditions.

As for LENA, the key components on the adult WCE pathway include detection of adult speech segments with a hidden-Markov model that uses so-called Minimum Duration Gaussian Mixture Models, application of a phone recognizer to the segments, and a linear mapping of the resulting vowel and consonant counts and speech duration measures to word counts as described in the previous subsection (Xu et al., 2008). Since the introduction of LENA, several studies have evaluated LENA WCE performance across a number of languages and participant populations (see Table 1). The major technical drawback of LENA is its reliance on the structure of American English phonology and lexicon in the WCE process. As LENA uses an acoustic phone model trained on En-

glish and a linear mapping from vowel and consonant counts to words, also optimized on English, the estimated word counts can be expected to be accurate only for languages that have the same ratio of vowels and consonants to words as the American English used in the training. In the reported literature, this problem is often masked by the use of Pearson's linear correlation between estimated and hand-annotated words as the primary performance metric to measure LENA reliability (e.g., Weisleder and Fernald, 2013; Soderstrom and Wittebolle, 2013; Canault et al., 2016; Gilkerson et al., 2015; Elo, 2016; Schwartz et al., 2017). Longer stretches of speech in any language also mean more words and subword units, so relatively high correlations between LENA output and manually annotated reference word counts have been reported in the literature for a variety of languages, as summarized in Table 1. However, the picture is very different when comparing the estimated counts N_{hypo} and true counts N_{true} with a measure that also considers the absolute counts, such as the mean absolute relative error used by Ziaei et al. (2016) in Eq. (1) or median absolute relative error in Eq. (2) used in the present study (more on evaluation in Section 4.1).

$$ERR_{\text{mean}}(\%) = \text{mean}\left(\frac{\left|N_{\text{hypo}} - N_{\text{true}}\right|}{N_{\text{true}}}\right) * 100 \quad (1)$$

$$ERR_{\text{median}}(\%) = \text{median}\left(\frac{\left|N_{\text{hypo}} - N_{\text{true}}\right|}{N_{\text{true}}}\right) * 100 \quad (2)$$

As an example, Elo (2016) reports a correlation of $r = 0.99$ between LENA and hand-coded word counts for several 1-hour segments from two Finnish children. At the same time, the estimated and actual word counts differ by 75.2%, most likely due to the highly different phonological and morphological structure of Finnish compared to English. This is not to say that LENA would not be applicable to languages different from English: a high correlation means that the relative word counts within the study population are still accurately measured, allowing word counts to be linked to other factors (e.g., developmental outcomes) with high validity. However, comparison of word counts across *different* participant populations is more problematic since LENA adult word counts are not guaranteed to correspond to actual words in a new language or in a substantially different dialect.[1]

In the present paper, we aim to remedy the problem of language-specificity by proposing a system that is always adapted to the target language using a small amount of orthographically transcribed speech. Since nearly all behavioral studies using LENA have checked the validity of the automated analyses in a given domain by comparing automated system outputs to manual annotations on a subset of the data (e.g., Soderstrom and Wittebolle, 2013; Weisleder and Fernald, 2013; Canault et al., 2016; Elo, 2016), the data transcribed for validity could also be used to adapt the WCE system to the domain in question (and one can still estimate validity by a cross-fold validation procedure on the same data). By implementing this type of low-resource adaptability, we aim for our system to be applicable to any language or use domain so long as the user is able to provide orthographic transcripts for roughly 30 min of audible adult *speech* (not to be confused with the total duration of annotated audio), ideally consisting of multiple different talkers and families. To validate our approach, we conduct experiments on six different corpora and use performance metrics that take into account the absolute accuracy of the estimator instead of measuring linear correlations. The overall purpose is to understand whether the adaptation approach is feasible with the amount of orthographically transcribed data that are manageable for language researchers to produce, and which technical components (SADs, syllabifiers) provide the best performance in WCE when overall accuracy and consistency across datasets are used as the primary criteria.

The rest of the paper is organized as follows: Section 2 introduces the proposed WCE system and its sub-components. Section 3 describes the data used in system training and in the experiments. Section 4 shows the results, and Section 5 discusses implications of the current work and how the system could (and should) be improved further in future work.

2. Methods

2.1. Overall WCE pipeline

A schematic view of the WCE pipeline,[2] largely based on the earlier work by Ziaei et al. (2016), is shown in Fig. 1. The system consists of five basic components (1) a speech activity detector (SAD), (2) a speech enhancement module (spectral subtraction), (3) automatic syllabification of speech input, (4) extraction of statistical descriptors from enhanced and syllabified signal representations, and (5) a linear mapping from features into corresponding word counts.

The guiding principles in the overall system design are robustness against signal conditions in daylong recordings and adaptability to new languages. The use of syllables as the primary feature for WCE is motivated by two primary reasons: First, signal-driven syllabification of speech can be viewed as a relatively language-independent process. This is due to an assumption that holds similarly across languages that syllabic nuclei are perceptually more "sonorous" than a preceding onset and (optionally) following coda (Whitney, 1874; de Saussure, 1916; Clements, 1990), where sonority is closely correlated with physical signal properties such as intensity or loudness (e.g., Price, 1980; Parker, 2002; see also Räsänen et al., 2018, for an overview). Any syllabification process that operationalizes these sonority fluctuations as generic computational transformations operating on the acoustic signal then remains largely language-independent. The second reason for using syllables is that the energetic nature of syllabic nuclei also makes them potentially robust against signal degradations such as additive noise. As long as the alternation between the less and more sonorous speech sounds is present in the signal representation, information on the number of syllables is also present.

Our system's adaptability to new languages is achieved by having only a small number of free parameters that depend on the language L in question: (a) a syllable detection threshold θ_L, (b) feature-specific coefficients $\boldsymbol{\beta}_L$ used in the linear mapping (highlighted in red in Fig 1), and (c) a correction coefficient α_L for limited recall of the used SAD. Given that a dataset with ~30 or more minutes of orthographically transcribed adult speech is available as training data, these parameters can be adapted to the language in question. A simple linear model is sufficient if we can assume that the word and syllable counts increase linearly with the duration of speech input at the time-scales of interest for WCE analysis (see also Xu et al., 2008; Ziaei et al., 2016). Parsimony in such a model also reduces the risk of having substantially different estimator behavior in different languages and conditions, as it mitigates the risks of overfitting the system to limited adaptation data or creating far more accurate models for high-source languages. Even if it excludes the potential benefits from more complicated nonlinear dependencies between signal descriptors and word counts, the previous WCE systems have also found linear model as suitable for the task (Xu et al., 2008; Ziaei et al., 2016).

[1] In practice, measuring linguistic exposure across different languages using absolute word counts is also problematic due to the large differences between languages at various levels of linguistic structure (see, e.g., Allen and Dench, 2015, for a discussion). However, this discussion is beyond the scope of the present study, where we simply aim to achieve similarly reliable WCE across languages.

[2] MATLAB implementation of the WCE system (source code + Linux/OSX Binaries for MATLAB Runtime Environment) are available at http://www.github.com/ACLEW/WCE_VM/, and also as integrated to the ACLEW virtual machine at http://www.github.com/SRVK/DiViMe/.

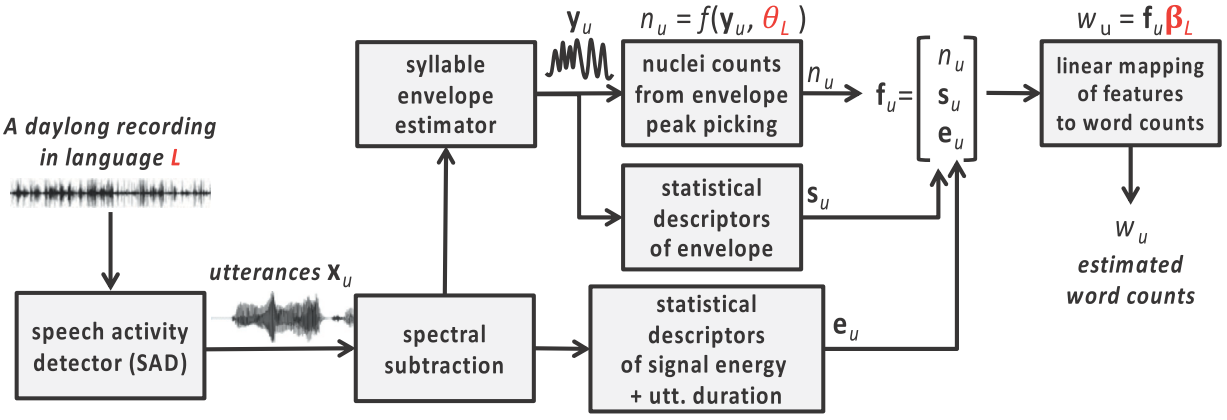

Fig. 1. Overall schematic view of the core WCE system. Input audio is first passed through a SAD that passes through detected speech segments, followed by speech enhancement with spectral subtraction. A syllabification algorithm is then used to calculate a "sonority envelope" $\mathbf{y}_u$ for each utterance u, from which syllable counts n_u are then obtained with peak picking. Utterance duration and a number of statistical descriptors of the enhanced audio and sonority envelope are then combined with the estimated syllable count to form a fixed-dimensional feature vector $\mathbf{f}_u$. Word count estimate of each utterance is finally obtained by applying a least-squares linear mapping to $\mathbf{f}_u$. Parameters θ_L and $\boldsymbol{\beta}_L$ shown in red font can be optimized separately for each language L. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

2.2. Processing steps in more detail

The processing of a (day)long recording starts with the detection of speech segments using a SAD.[3] In this work, we compare three alternative methods for the task: Threshold-optimized Combo-SAD ("TO-Combo-SAD; Sadjadi and Hansen, 2013; Ziaei et al., 2014) as used in Ziaei et al.'s WCE system (2016), SAD from the widely used OpenSMILE toolbox (Eyben et al., 2013a, 2013b), and so-called 'Noisemes' SAD, which is based on recognition of several classes of environmental sounds (Wang et al., 2016), all described in more detail in the next subsection.

All segments classified as speech by the SAD are subsequently processed by a speech enhancement algorithm. Our system uses spectral subtraction (Berouti et al., 1979), which Ziaei et al. (2016) found to be superior to several other methods in their comparisons on the Prof-Life-Log WCE experiments. In the present system, spectral subtraction is carried out using the noise power spectral density estimation algorithm by Martin (2001), as implemented in the VoiceBox toolbox,[4] where noise estimation is performed directly from the SAD output segments without having to separately specify non-speech regions. This simplifies the pipeline, as a SAD may not always reliably differentiate between speech and non-speech content (as will be seen in the experiments below).

In the syllable envelope estimator stage, we compare two unsupervised syllabifiers: one by Wang and Narayanan (2007) and one by Räsänen et al. (2018). In addition, we investigate a supervised neural network-based syllabifier based on an initial concept in Landsiedel et al. (2011). In all three, the enhanced acoustic waveform corresponding to a SAD output segment ("utterance") u is transformed into a unidimensional signal $\mathbf{y}_u \in [0, 1]$ at a 100-Hz sampling rate. Each sample in $\mathbf{y}$ represents either "sonority" of the speech input at that instant (for unsupervised estimators) or pseudo-probability of a syllable nucleus at the given time (for the supervised estimator). As a result, local peaks in $\mathbf{y}$ are assumed to correspond to syllabic nuclei.

In the feature extraction stage, the number of syllable nuclei n_u is first extracted from the syllable envelope $\mathbf{y}_u$. This is performed using a simple peak-picking algorithm that looks for local maxima with amplitude differences of at least θ_L with respect to the previous local minimum. The threshold parameter θ_L is optimized separately for each language L (see Section 2.2). In addition to n_u, the mean and standard deviation (SD) of the sonority envelope across the entire utterance are extracted as syllabic features $\mathbf{s}_u$. The mean and SD of signal power and overall SAD segment duration are also extracted as signal-level energy features $\mathbf{e}_u$. Even though the mean and SD features do not accumulate over time, initial experiments suggested that they allow automatic fine-tuning of predictions based on overall signal dynamics in the utterances.

In the final stage, all features n_u, $\mathbf{s}_u$, and $\mathbf{e}_u$ are concatenated into an utterance-level feature vector $\mathbf{f}_u$, and a linear mapping $w_u = \mathbf{f}_u\boldsymbol{\beta}_L$ to the corresponding word count estimate w_u is carried out. Similarly to θ_L, the mapping parameters $\boldsymbol{\beta}_L$ are separately optimized for each language.

2.3. Adapting the system to new languages

In order to adapt the system to a new language L, syllable detection threshold θ_L and linear mapping parameters $\boldsymbol{\beta}_L$ are estimated from utterances $\boldsymbol{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$ for which the corresponding word counts $\boldsymbol{w} = [w_1, w_2, \ldots, w_n]$ are known. The parameters are optimized to minimize WCE RMSE error on the provided training data. This is achieved by first performing syllabification of training utterances at various threshold values $\theta \in [0.0001, 1]$ with small increments and measuring the linear correlation r between the resulting syllable counts and ground-truth word counts across all the utterances. The threshold θ_L with the highest linear correlation is then chosen, and the corresponding syllable counts n_u are added to the utterance-level feature vectors $\boldsymbol{F} = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n]^T$ along with the other features. Ordinary least squares linear regression is then carried out to solve $\boldsymbol{\beta}_L$ from $\boldsymbol{w} = \mathbf{F}\boldsymbol{\beta}_L$.

In order to compute word count estimates over longer time-scales than individual SAD segments, a correction based on the expected recall of the SAD needs to be taken into account. In the experiments described in Section 4, SAD is first used to split the adaptation recordings into utterance-like chunks u, and then the proportion $\alpha_L \in [0, 1]$ of words ending up in the SAD outputs (see Section 4 for details) with respect to the total number of words in the adaptation data is measured. All aggregate word count estimates are then divided by α_L to account for the limited recall of the SAD.

2.4. Compared speech activity detectors

Three different SADs were compared in the experiments. The first two, TO-Combo-SAD and OpenSMILE SAD, are well established and

[3] We will refer to voice activity detectors (VADs) and speech activity detectors (SADs) simply as SADs.

[4] http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html, by Mike Brooks.

have been previously tested in a variety of contexts. The third one, Noisemes SAD, differs from the other two by attempting to model non-speech categories in more detail instead of directly attempting binary classification between speech and non-speech.

2.4.1. TO-Combo-SAD

TO-Combo-SAD (Sadjadi and Hansen, 2013; Ziaei et al., 2014) is based on five signal features (harmonicity, clarity, prediction gain, periodicity, and spectral flux) that are linearly mapped into a 1-D representation using PCA, and then clustered into two categories using a 2-component Gaussian mixture model (GMM) based on the data from the analyzed segment. In the basic Combo-SAD, the GMM component with the higher mean is then considered to be speech and the other component non-speech. The final frame-level decisions are made based on the component posteriors after weighing them with factor w (and 1-w) that is a hyperparameter of the algorithm. In the threshold-optimized version used in this paper, the higher component mean has to be equal to or higher than the mean of 1–D projections of all mean vectors from a 256-component GMM, where this larger GMM that has been pre-trained on a large amount of labeled speech data using the same 5-dimensional features. If this condition is not satisfied (i.e., neither cluster resembles typical speech), the pre-trained GMM is used as a model of speech instead. As a result, TO-Combo-SAD is capable of handling audio data with highly unbalanced distributions of speech and non-speech content, as demonstrated with Apollo space mission data (Ziaei et al., 2014) and, in WCE, with the Prof-Life-Log data (Ziaei et al., 2016). In the present experiments, we use this "threshold-optimized" ("TO") Combo-SAD with its default parameters, as kindly provided by the original authors.

2.4.2. OpenSMILE SAD

OpenSMILE SAD (Eyben et al., 2013a, 2013b) is included as the second SAD alternative, since the OpenSMILE toolkit is widely used for various speech processing applications and is freely available for non-commercial use. The SAD of the toolkit uses a Long Short-Term Memory (LSTM) neural network model with cepstral coefficients computed from RASTA-PLP (Hermansky and Morgan, 1994) and their first and second order derivatives. During use, network outputs (−1 for non-speech, 1 for speech) are thresholded to make a binary speech/non-speech decision for each frame. In Eyben et al. (2013a), the network was trained using American English speech corpora of conversational (Buckeye; Pitt et al., 2005) and read speech (TIMIT; Garofolo et al., 1990) and using synthetic additive noise for improved noise robustness. However, the public version available in the OpenSMILE toolbox uses a more limited training dataset that is not separately specified (see OpenSMILE documentation). Default hyperparameters of the tool (OpenSMILE version 2.1.0) were used in the experiments.

2.4.3. Noisemes SAD

Noisemes SAD (Wang et al., 2016) was chosen as the third alternative SAD, since it represents a somewhat different approach to the speech detection problem than the previous two: It is, in fact, a 17-class environmental sound ("noiseme") classifier with two categories for speech and 15 categories for other sound types, such as music, singing, cheering, and mumbling. Since in WCE we want to distinguish comprehensible speech from other vocalizations, this type of multi-class modeling may be beneficial. Technically Noisemes SAD is based on 6669 low-level signal descriptors extracted using the OpenSMILE toolkit that have been compressed to 50-dimensional features using PCA,[5] and fed into a one-layer Bidirectional Long Short-Term Memory (BLSTM) network. The model has been trained on 10 h of web video data from Strassel et al. (2012). To use it as a SAD in our experiments, posteriors for "speech" and "speech non-English" classes were summed together and all frames where this combination class was higher than the other 15 categories were considered to be speech.

[5] Note that the original method in Wang et al. (2016) used 983 features selected using information gain criterion, but we used an updated version from authors Wang and Metze in this paper.

2.5. Compared syllabifiers

The basic idea of tracking sonority fluctuations in speech has given rise to several automatic syllabification algorithms proposed in the existing literature. Even though there is variation in the exact framing of the methods, basing syllable detection on, e.g., amplitude or energy envelopes, loudness contours, or other similar 1-D representations derived from the signal (e.g., Mermelstein, 1975; Morgan and Fosler-Lussier, 1998; Villing et al., 2004; Wang and Narayanan, 2007; Obin et al., 2013), nearly all of the methods are ultimately based on tracking of the approx. 3–8 Hz amplitude modulations in the speech signal that go hand-in-hand with the temporal alternation between vocalic and consonantal speech sounds: the syllabic rhythm.

In the present work, three alternative syllable envelope estimators were compared for WCE: (1) thetaSeg algorithm by Räsänen et al. (2018), originally designed for perceptually motivated syllable segmentation from speech, (2) syllable envelope-estimator module from the speech-rate estimator by Wang and Narayanan (2007), and (3) a bi-directional Long Short-Term Memory (BLSTM)-based syllabification algorithm based on the initial version described in Landsiedel et al. (2011). While the first two alternatives (thetaSeg and WN) are unsupervised methods making use of heuristic signal processing operations, the BLSTM is directly trained for nucleus detection in a supervised manner. All three methods are detailed below.

2.5.1. thetaSeg

thetaSeg (Fig. 2, top; Räsänen et al., 2018) is a straightforward mechanistic model of oscillatory entrainment of the auditory cortex to rhythmic fluctuations in speech input (approx. 4–7 Hz; so-called "theta-range" oscillations), approximating the perception of sonority fluctuations in speech. In thetaSeg, the incoming signal is first fed through a 20-channel Gammatone filterbank with center frequencies logarithmically spaced between 50 and 7500 Hz, followed by downsampling of the amplitude envelopes to 1000 Hz. The resulting envelopes are then used to drive a bank of harmonic damped oscillators (2nd order electronic resonators with shared parameters), one oscillator for each frequency band. For each sample, amplitudes of the eight highest-amplitude oscillators are combined non-linearly by calculating the sum of logarithmic amplitudes of the oscillators. As a result, a time-series called "sonority envelope" is obtained, where harmonic and high-energy in-phase excitation on multiple frequency bands is reflected as high amplitude values, whereas low-energy and/or incoherent excitation will result in smaller values. The damping and center frequency parameters of the thetaSeg were optimized in Räsänen et al. (2018) for maximal syllable segmentation performance across English, Finnish, and Estonian conversational speech. The resulting damping factor $Q = 0.6$ and center frequency $cf = 8$ Hz are also used in the present paper.

2.5.2. WN

WN (Wang and Narayanan, 2007; Fig. 2, middle) is an algorithm originally developed for speaking rate estimation from conversational speech, being an improved modification of the mrate-algorithm proposed by Morgan and Fosler-Lussier (1998). WN is also used by Ziaei et al. (2016) in their WCE system. In WN, the signal is first divided into 19 frequency bands, followed by downsampling to 100 Hz, and selection of the 12 most energetic sub-bands. For each sub-band, the envelopes are low-pass filtered with a Gaussian-shaped kernel, followed by computation of temporal within-band correlations up to lag of $K = 11$ frames. The resulting band-specific signals are then combined through multiplication (∼cross-band correlation), and smoothed again

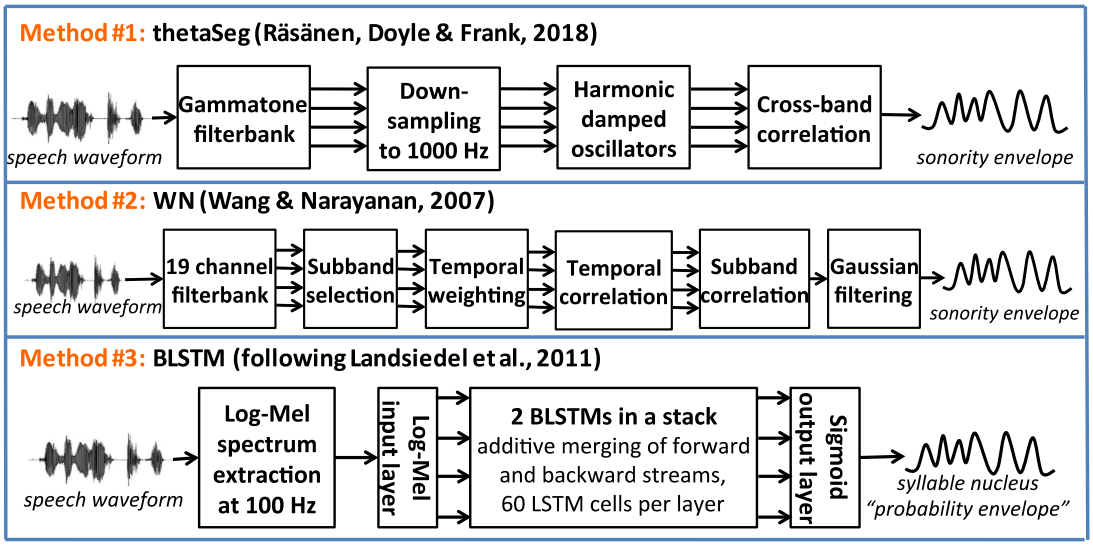

Fig. 2. Block schematics of the syllable envelope estimators compared in the present study.

in time using a different Gaussian kernel. As a result, a one-dimensional sonority-like envelope is obtained, in which peaks are assumed to correspond to syllabic nuclei.

Our experiments used a MATLAB implementation of WN that is described in Räsänen et al. (2018). In that version, the envelope estimation stage is identical to the original one described in Wang and Narayanan (2007) except that a Gammatone filterbank was used instead of the original second-order Butterworth bandpass filters for the frequency analysis. All hyperparameters (number of frequency bands and sub-bands, Gaussian kernel sizes etc.) were taken from the original paper, where they were optimized for the conversational Switchboard corpus (Godfrey et al., 1992) using a Monte Carlo optimization scheme. The original speech rate estimator also uses pitch tracking to prune out unvoiced nucleus candidates. However, robust F0 estimation with a fixed set of hyperparameters was found to be problematic across the variety of signal conditions encountered in our daylong recordings. Therefore we only used the envelope estimation stage of the WN, and the envelope was used as an input to the same peak picking algorithm used by all three syllabifiers (as described in Section 2.2).

2.5.3. BLSTM syllabifier algorithm

BLSTM syllabifier algorithm (Fig. 2, bottom) was developed based on the initial work by Landsiedel et al. (2011) who tested BLSTM-based syllabification on English from TIMIT and Switchboard corpora. However, instead of a higher-dimensional set of features used in the original paper, inputs to our model are mean and variance normalized 24-channel log-Mel spectra (25-ms frames, 10-ms frame shift). We also doubled the number of units in hidden layers to support representation learning from the spectral input. As a result, the network uses two bi-directional layers with 60 LSTM cells in each forward and backward layer, and where forward and backward layer LSTM cell activations are combined through addition. Sigmoid activation functions are used for each LSTM cell. After merging the final BLSTM layers, there is a fully-connected sigmoid layer with one node that converts the BLSTM activations into syllable nucleus probabilities, one value for each input frame.

Training of the BLSTM was carried out using syllable-annotated data from several different languages described in Section 3.1. Target outputs for network training consist of 1-D time-series that are otherwise zero except for Gaussian-shaped kernels centered on manually annotated syllable nuclei. More specifically, for each phone that is also a syllabic nucleus, a Gaussian kernel with a maximum value of one is added to the position corresponding to the center of the phone. The standard deviation of the Gaussian is set to be the corresponding phone duration, divided by 3.5. Any values larger than one, basically due to temporally overlapping Gaussians, are clipped to have a value of 1. As a result, the target signal can be interpreted as a pseudo-probability for the presence of a syllabic nucleus in each position (see also Landsiedel et al., 2011). The use of Gaussians instead of binary targets accounts for the various sources of uncertainty in determining the accurate position and duration of a syllabic nucleus, including coarticulatory effects and annotator variability, and even conceptual problems in defining the exact onset and offset of a syllabic nucleus.

In our experiments, we explore four alternative training strategies for the BLSTMs: (1) clean training without dropout, (2) clean training with 50% dropout in the hidden layers, and additive noise and varying channel augmented training (3) with and (4) without dropout (50%). Noise and channel augmentation were carried out by creating two additional copies of each clean training signal. For each copy, additive noise was added at SNR sampled uniformly and randomly from [−10, 40] dB. The additive noise signals consisted of randomly sampled extracts from ACLEW starter set (Bergelson et al., 2017a) consisting of infant daylong recordings from various language environments, none of the data drawn from the test participants of our experiments. Varying channel characteristics were simulated by convolving the noised speech samples with FIR filters of 20 coefficients ($fs = 16$ kHz) randomly sampled from a normal distribution with zero mean and unit variance. The resulting signals were scaled to have a maximum amplitude of 1. The motivation for the data augmentation was to explore whether this type of approach improves syllabification performance also in case of largely unconstrained auditory environments present in our recordings, and also how augmentation compares with the effects of dropout training in our application.

3. Data

Two separate sets of data were used in the development and testing of the WCE pipeline: one set of corpora for training the BLSTM syllabifier, and another set of corpora of daylong child recordings for testing of the WCE system.

3.1. BLSTM syllabifier training data

The BLSTM syllabifier was trained on data from four corpora that have both syllable- and phone-level annotations available: the Phonetic Corpus of Estonian Spontaneous Speech ("EstPhon"; Lippus et al., 2013), the Korean Corpus of Spontaneous Speech (Yun et al., 2015),

Table 2
Corpora used in the experiments for WCE evaluation. *Audio total* = total amount audio annotated for verbal activity; *speech total* = duration of all utterances in the annotated audio; *adult speech total* = total duration of utterances from male or female adults that contain at least one unambiguously transcribed word. *Min* = minutes.

ID	Corpus name	Language	*Subjects (N)*	Audio total (h)	Speech total (min)	Adult speech total (min)	Audio per subject (min; avg.)
BER	Bergelson	US English	10	5.0	116.7	50.7	30.0
CAS	Casillas	Tseltal	10	7.5	212.0	100.8	45.0
L05	Language 0–5	UK English	10	5.0	95.9	39.1	30.0
ROS	Rosemberg	Arg. Spanish	10	5.0	149.3	70.3	30.0
MCD	McDivitt+	Can. English	8	4.5	80.9	44.0	33.8
WAR	Warlaumont	US English	10	5.0	100.3	39.6	30.0
		Total	58	32.0	755.1	344.5	

the Brent corpus of American English infant-directed speech (Brent and Siskind, 2001), and the C-PROM corpus of spoken French (Avanzi et al., 2010). Together these corpora cover four different languages, several speaking styles, and a range of recording conditions from speakers of both genders and across a variety of ages.

From EstPhon we used the studio section of the corpus, which includes several spontaneous dialogues with pairs of male and female talkers, totaling up to 10,158 utterances in high signal quality (5.2 h of audio). The Korean data consists of dialogues between talkers of various ages (from teenagers to speakers in their 40 s) and both genders. Since this corpus is much larger than the other three, we randomly sampled a subset of 12,000 utterances (5.0 h) for training. As for C-PROM, we used the entire corpus consisting of 24 multi-minute recordings of various regional varieties of French from several discourse genres, totaling 1.2 h of data. Finally, we used the so-called Large-Brent subset of the Brent corpus forced-aligned for words and phones by Rytting et al. (2010), for which automatic syllabification of the resulting phone transcripts was carried out using tsylb2-algorithm (Fisher, 1996), as described in Räsänen et al. (2018). This subset of Brent corresponds to 1.9 h of speech and 6253 utterances. After combining all the four corpora, the training data consisted of 13.3 h of audio with 265,089 syllables. In data augmentation experiments, this was tripled to 40 h, as described in Section 2.5.

3.2. Evaluation data

The data for WCE system evaluation comes from six different corpora of child daylong recordings that have been pooled together, sampled, and annotated as part of the ACLEW project (Bergelson et al., 2017b). These include the Bergelson corpus ("BER") from US English families from New York area (Bergelson, 2016), the LuCiD Language 0–5 corpus ("L05") consisting of English-speaking families from Northwest England (Rowland et al., 2018), the Casillas corpus ("CAS") of Tseltal-speaking families from a rural Mayan community in Southern Mexico (Casillas et al., 2017), the McDivitt and Winnipeg corpora (so-called McDivitt+; here "MCD") of Canadian English families (McDivitt and Soderstrom, 2016), the Warlaumont corpus ("WAR") of US English from Merced, California (Warlaumont et al., 2016), and the Rosemberg corpus ("ROS") of Argentinian Spanish families from Buenos Aires metropolitan area (Rosemberg et al., 2015). Some recordings in BER, and all recordings in CAS, MCD, and WAR are available from HomeBank repository (VanDam et al., 2016).

Key properties of these corpora are summarized in Table 2. Each corpus consists of daylong (4–16 h) at-home recordings; each corpus samples from a unique community, with language varying across corpora and socioeconomic environment varying both within and across corpora. In each recording, the target child ("participant") wears a mobile recorder in a special vest throughout a normal day. BER, MCD, L05, and WAR recordings were collected with the LENA recorder, while CAS was recorded with Olympus WS-382 or WS-852, and ROS was recorded with a mix of Olympus, Panasonic, Sony, and LENA recorders. All the recorders have high-quality microphones on speech frequency band. All data were recorded at a 16-kHz sampling rate or higher at 16 bits, and converted to .mp3 for cloud storage on Databrary (https://nyu.databrary.org/). All data were resampled to 16 kHz before further processing. Due to the unconstrained nature of the recordings, they contain both near- and far-field speech in various ambient environments and at highly varying SNRs. The approximate[6] average speech SNRs for different corpora are BER 2.1 dB, CAS –0.5 dB, L05 3.6 dB, ROS –2.6 dB, MCD 0.8 dB, and WAR 2.4 dB.

Out of the 220 of recorded participants, daylong recordings from 10 infants from each corpus were chosen for manual annotation, selected to represent a diversity of ages (0–36 months) and socio-economic contexts. From those daylong files, fifteen 2-minute non-overlapping segments were randomly sampled from the entire daylong timeline for manual annotation, corresponding to approximately 10 min of annotated speech per subject. The only exception to this is the CAS corpus, which consists of nine randomly sampled 5-min segments for each of the 10 children. It also contains 50% more annotated audio than the others, since all of its annotations were carried out before determining the final sampling protocol for the rest of the corpora. One MCD subject from a French-speaking family was excluded from the experiments, as the other subjects were from English-speaking families. Due to a sampling error, one of the remaining participants was sampled twice.

All sampled 2- and 5-min segments were annotated for all hearable utterance boundaries, speaker ID, addressee information (child vs. adult-directed speech), and vocal maturity of child vocalizations (canonical/non-canonical babbling, single-, or multiword utterances), and all adult speech was transcribed. All annotations followed a shared annotation protocol developed in the ACLEW project for the present type of daylong data (Casillas et al., 2017a, 2017b). Each corpus was annotated by (or with) someone proficient in the language in question. To ensure standardization in annotation, all annotators passed a test against a (separate) reference gold standard annotation before annotating the data here. Annotators were trained to transcribe speech corresponding to what was actually said instead of the canonical lexical forms (e.g., '*wanna*', not '*want to*').

For the WCE experiments, reference word counts were extracted from the orthographic transcripts of the utterances. First, all non-lexical transcript entries such as markers for incomprehensible speech, non-linguistic communicative sounds, and all paralinguistic markers were discarded. In addition, for ready comparison to LENA, all transcribed words from non-adult speakers were discarded, even though the present WCE pipeline does not yet have a mechanism for separating speaker identities. As a result, only unambiguously transcribed word forms from adult talkers remain in the final gold standard dataset here. Every remaining orthographic entry separated by a whitespace was then considered as a word for adapting and testing of the WCE system.

[6] Noise power was estimated as the mean power of non-speech frames within a 5-s window centered around each annotated speech frame (defaulting to average signal noise power if no non-speech frames were within that window). Speech power was estimated from speech frames by assuming non-coherent additivity of speech and the noise estimate for the given frame.

4. Experiments and results

4.1. Experimental setup and evaluation

The purpose of the experiments was to evaluate our WCE pipeline across the different corpora, and to compare the alternative syllabification and SAD algorithms described in Section 2. Leave-one-subject-out (LOSO) validation was used to perform WCE on the data described in Section 3.2, always adapting the WCE system on all but one of the subjects, and then testing WCE performance on the held out subject. Adaptation and testing were carried out separately for each of the six corpora.

In addition to the three SADs (TO-Combo-SAD, OpenSMILE, Noisemes), we evaluated WCE performance with an ideal segmentation based on the utterance boundaries extracted from manual annotation. We also included a baseline "fixed-frame" segmentation condition where the audio signals were simply divided into fixed 5-s non-overlapping windows without any knowledge of the underlying signal contents, thereby passing all speech and non-speech audio content to the syllabification stage. For all these conditions, the six syllabifiers (thetaSeg, WN, and the four BLSTM variants) were compared against each other. In addition, a baseline system using only speech duration as the feature for least-squares linear regression was included.

In order to perform WCE adaptation, SAD was always first applied to the training data. Since the orthographic transcripts were aligned at the utterance (but not the word) level, the following procedure was used to assign transcribed words to SAD output segments: First, transcribed words of an utterance were assumed to be uniformly spaced across the entire duration of the utterance, where word duration was assumed to be directly proportional to the number of characters in the word. All transcribed words overlapping with the given SAD output segment were then assigned to it. Finally, all SAD segments $\mathbf{x}_u$ with their corresponding number of words w_u were considered as inputs to the optimization of the linear mapping. Correction factor for SAD recall was measured on the training data by dividing the number of words assigned to SAD outputs by the total number of words in the training data.

During testing, the time-scale of interest was all the audio data from the given subject s, corresponding to 30–45 min of total audio and approximately 4–10 min of adult speech per subject (Table 2). Estimated word counts from all SAD segments of a test subject were summed together to obtain $N_{s,\text{est}}$ and compared against the corresponding total number of annotated words $N_{s,\text{true}}$ in order to derive subject-level deviation ERR_s (%) between true and estimated word counts:

$$\text{ERR}_s = \frac{|N_{s,\text{est}} - N_{s,\text{true}}|}{N_{s,\text{true}}} * 100 \quad (3)$$

However, one challenge in this type of evaluation is that not all samples in our corpora necessarily contain adult speech ($N_{s,\text{true}} = 0$). One option is to simply ignore these samples, but that would bias the evaluation towards "easier" cases where a larger proportion of the signal timeline is covered by target speech (a problem that would also apply for weighted averages based on the reference counts). Another option—the one we adopted here—is to replace zeros with ones in the denominator in order to get a finite measure for each sample. In addition, we observed that the error distribution across subjects tends to be non-Gaussian with typically one or two outliers in nearly all corpora, thereby also increasing the mean of errors substantially above of the values typical to the overall pool of subjects (see also Table 1). As a practical compromise, we adopt the use of $\text{ERR}_{\text{median}}$ in Eq. (2) as the primary performance metric across all the subjects in a corpus, as it takes data from all subjects into account without assuming normality of the error distribution.

As for benchmarking against LENA, the English corpora BER, L05, MCD, and WAR were fully collected with LENA, and therefore LENA automated analysis outputs were available for comparison. Since LENA output consists of estimated adult word counts per each automatically detected conversational turn, these conversational turns had to be aligned with the 2-min segments sampled for manual annotation. In practice, if x% of a LENA conversational turn overlapped with a segment, x% of the LENA word counts from that turn were added to the total LENA word count estimate for that segment. Note that the proportion of partially overlapping turns is only 14% of all conversational turns. In addition, any assignment errors (i.e., too few or many words added to the given segment) resulting from this type of alignment procedure are independent of each other and have an expected value of zero, and hence the actual error also approaches zero when all the 15 segments from a subject are pooled together to get subject-level reference counts. Therefore, the performance figures reported for LENA below can be considered representative, even if exact alignments between LENA outputs and the current audio samples were not available. Since our proposed system uses adaptation-based scaling of signal and syllable features into word counts in each corpus, we also experimented with corpus-dependent least-squares linear scaling of LENA word counts using the same LOSO protocol. However, this did not lead to consistent performance improvements[7] on the held-out data across the four English corpora (for which LENA outputs were available), and therefore we report LENA performance based on default LENA output.

4.2. Main results

Fig. 3 shows the results from the main WCE performance evaluations for each SAD, syllabifier, and corpus. In addition Fig. 4 shows the performance of the BLSTM syllabifier (with augmentation and dropout) on the six different corpora as a function of observed speech when ideal utterance segmentation is used.

In case of ideal utterance segmentation (Fig. 3, top), all compared syllabifiers perform relatively well with the median estimation error being below 10% in nearly all cases. The duration baseline also reaches relatively good performance levels, but is still on average worse than the syllabifier-based approaches. Fig. 4 also demonstrates how the estimation error decreases practically linearly in the log/log-domain when more speech is observed: Even if the accuracy at the level of individual utterances is not high (around $\text{ERR}_{\text{median}} = 40$–$60\%$), the estimate becomes gradually more accurate over time. In general, it appears that approximately 100 s of adult speech would be sufficient for approx. ~10% relative error in word counts independently of the language, assuming that the SAD used is perfectly accurate. However, one can also see from the zoomed-in region of Fig. 4 that the performance improvements on L05 corpus appear to start to saturate at approx. 10% relative error level after one minute of observed speech data. This demonstrates in practice the fact that there is no guarantee that the system, when optimized to minimize the WCE error across all the adaptation data, would be fully unbiased on any individual subject.

As can be expected, overall performance is notably lower in the more realistic use case with actual SADs (Fig. 3, panels 2–4). In addition, clear differences between the corpora and SADs can be observed. For instance, the use of TO-Combo-SAD leads to relatively good performance across the board, except for WAR corpus where the errors are three-fold compared to the other corpora. In contrast, Noisemes SAD does slightly better than TO-Combo-SAD on WAR, but performs poorly on ROS and MCD. OpenSMILE SAD has very good performance on some of the corpora and also the best performance of all on WAR, even though performance on BER, MCD, and L05 is slightly worse than with TO-Combo-SAD. Interestingly, fixed-frame segmentation without any speech detection front-end outperforms the Noisemes SAD when the BLSTM syllabifier is used. The overall pattern of results suggests that none of the tested SADs are well-rounded performers, and suitability of the SAD for a given recording environment has an effect on overall system performance. OpenSMILE

[7] Performance improved slightly on BER and MCD whereas it got worse on L05 and WAR. The average effect of linear scaling across the four corpora was 0.34% absolute decrease in median absolute relative error rate.

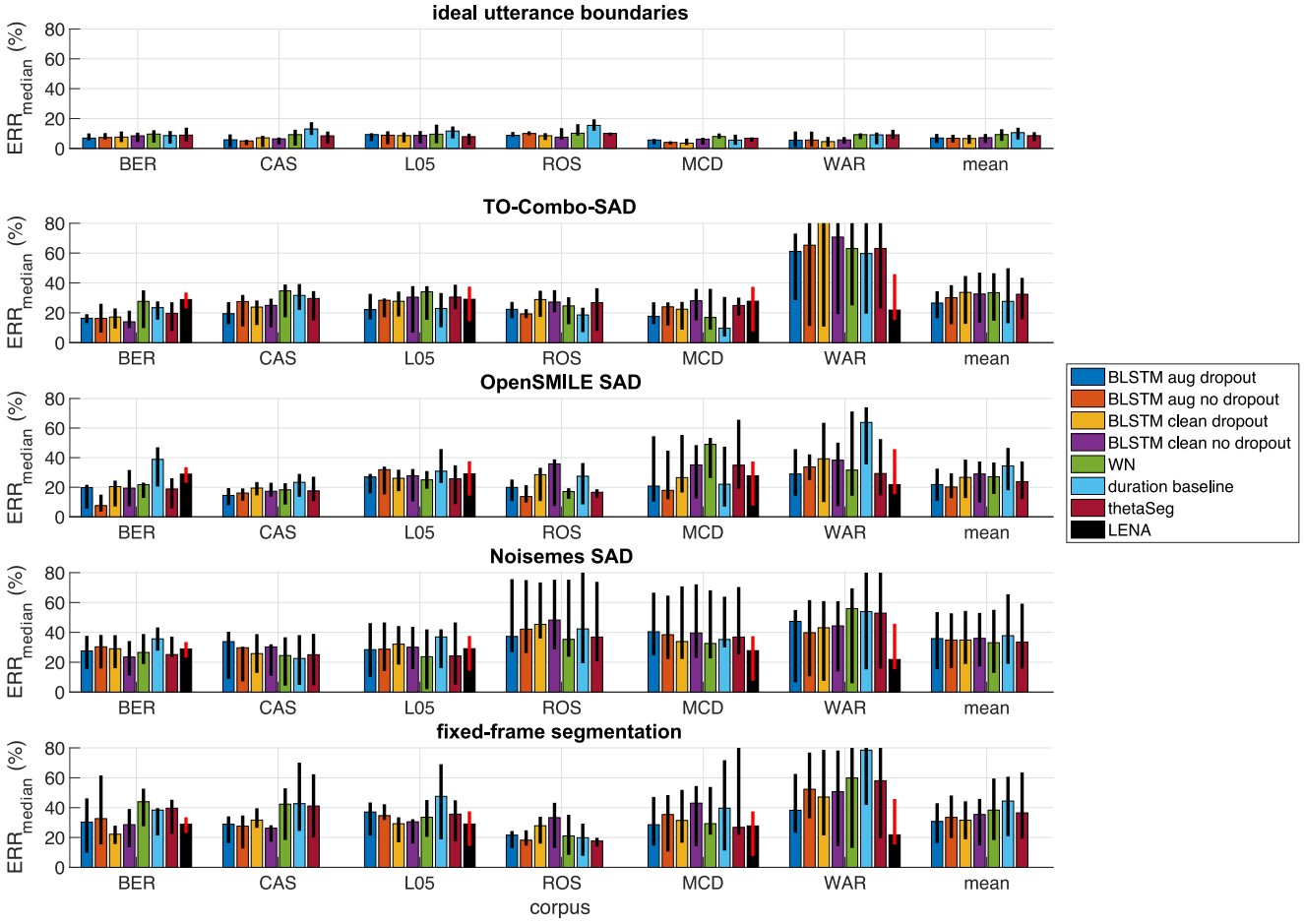

Fig. 3. Results from the main experiments. Each panel shows the WCE error rate ERR_{median} for each corpus and syllabifier when using a specific SAD. The mean across the corpora is also shown for each syllabifier. Top panel: TO-Combo-SAD. Second panel: OpenSMILE SAD. Third panel: Noisemes SAD. Bottom panel: fixed 5-s segments. LENA reference performance is shown with black bars where available. 3rd and 7th deciles are shown with vertical bars. Note that LENA performance is only shown for reference and does not depend on the SADs tested in the present experiments.

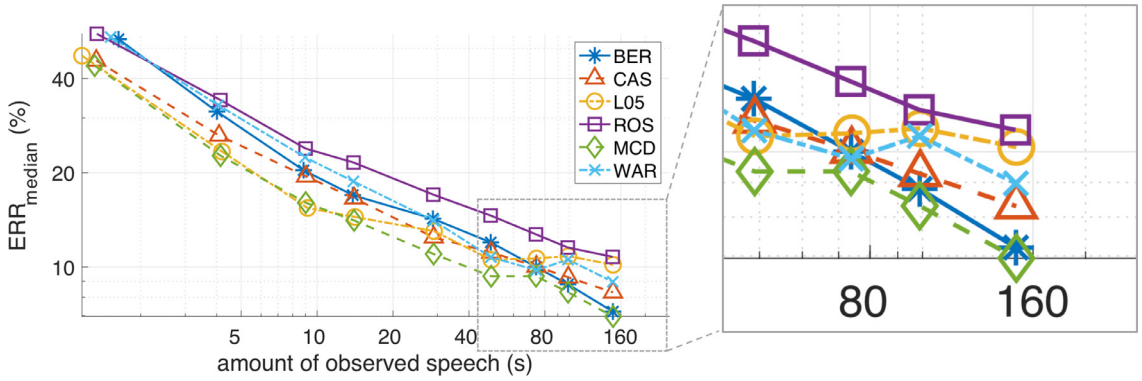

Fig. 4. Performance as a function of the duration of observed speech when using ideal segmentation into utterances from adult speakers. Results are shown for the BLSTM syllabifier trained with data augmentation and dropout.

SAD has the best average performance of all the alternatives whereas Noisemes SAD is clearly not as suitable for the present task. To-Combo-SAD would otherwise be on par with OpenSMILE, but the problems with WAR cause its mean performance to deteriorate substantially.

As for the compared syllabifiers, the BLSTM generally outperforms the unsupervised methods thetaSeg and WN when either of the two well-performing SADs or fixed-frame segmentation is used, especially when training data augmentation has been used. WN and thetaSeg perform approximately at the same level with the non-augmented BLSTMs with or without dropout with some variation across corpora and SADs, but fall behind the augmented BLSTMs in overall accuracy and consistency. This demonstrates that the multilanguage training of the BLSTM indeed works so that the models generalize to novel languages, and that the BLSTM is more tolerant against non-speech noise in the signals than the unsupervised methods. The results also suggest that training data augmentation and dropout training are both useful in the task despite the hundreds of thousands of syllable exemplars available in the training data.

Table 3
Average SAD performance for different corpora and SADs compared, metrics averaged across all LOSO test sets on the given corpus. P = precision, R = recall, F = F-score.

	BER			*CAS*			*L05*			*ROS*			*MCD*			*WAR*			*mean*		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
TO-Combo-SAD	0.34	0.45	0.38	0.58	0.59	**0.58**	0.29	0.48	**0.36**	0.53	0.51	**0.51**	0.26	0.43	**0.34**	0.29	0.44	**0.32**	0.38	0.48	**0.41**
OpenSMILE SAD	0.26	**0.86**	**0.40**	0.39	**0.88**	0.53	0.20	**0.88**	0.31	0.36	**0.80**	0.49	0.14	**0.85**	0.27	0.19	**0.84**	0.29	0.26	**0.85**	0.38
Noisemes SAD	**0.51**	0.26	0.34	**0.66**	0.22	0.32	**0.39**	0.30	0.32	**0.56**	0.14	0.20	**0.30**	0.21	0.25	**0.38**	0.31	0.31	**0.47**	0.24	0.29
mean	0.37	0.52	0.37	0.54	0.56	0.48	0.29	0.55	0.33	0.48	0.48	0.40	0.23	0.50	0.29	0.29	0.53	0.30	0.37	0.53	0.36

Looking at the different languages tested, the proposed WCE system, especially the augmented BLSTM with dropout and TO-Combo-SAD, seems to reach similar performance levels in English (e.g., BER, L05, MCD), Spanish (ROS), and Tseltal (CAS). When the same syllabifier is paired with the OpenSMILE SAD, the resulting performance is also independent of the language in question. This demonstrates that the language as such is not a key factor in determining system performance, and that the adaptation procedure seems to work. However, it also seems that one of the English corpora, WAR, is more challenging than the others as indicated by higher error rates in comparison. To understand why this is the case, we manually investigated the properties of WAR and compared it to the other English corpora. Even though it is difficult to determine the exact source of the differences, WAR was found to have the least adult speech among the corpora, several subjects having extremely few unambiguously transcribed adult words across all the audio for the subject. In contrast, the proportion of infant's own vocalizations, the amount of background electronic speech (e.g., TV or radio, not transcribed for words), and the amount of adult singing was found to be high for several subjects in WAR, potentially causing problems for the SADs and for the WCE that is currently unable to distinguish different sources of speech. Since WAR still has similar performance to others in case of ideal utterance segmentation, this suggests that the errors are related to the lack of sufficiently well-performing mechanisms for detecting adult speech from the audio recordings. We also verified that the WCE performance of the BLSTM system did not correlate with the average SNR of each corpus ($p > 0.05$, Pearson correlation), likely since the SNR differences between the corpora are small (Section 3.2).

Finally, comparison to LENA shows that the TO-Combo-SAD + BLSTM system outperforms LENA on a number of varieties of English (American in BER, British in L05, and Canadian in MCD), even though LENA has been optimized for American English. Only in the case of WAR, performance is worse in the present system than in LENA, again suggesting that the present system's speaker attribution mechanisms are poorer than those of LENA, at least for the present task. Unfortunately, no LENA output data were available for the Spanish or Tseltal corpora to enable comparison on the non-English languages (but see Fig. 8 in Section 5).

4.3. SAD impact on WCE performance

The main results in Fig. 3 suggest that SAD performance on the test data might be one key factor behind the differences in the WCE performance. To study this further, Table 3 shows the performance of the three SADs on each of the corpora, declaring as "speech" the adult utterances that contain one or more unambiguously transcribed words and all other sections mapping to "non-speech" or "silence".[8] The table reports *precision* (the proportion of speech frames hypothesized to be speech actually being speech), *recall* (the proportion of all true speech frames detected), and *F-score* (the harmonic mean of precision and recall).

As the data shows, all three SADs have very different operating points on the daylong infant data. While TO-Combo-SAD has a more balanced precision and recall (though still having problems with precision on L05, MCD, and WAR), OpenSMILE SAD reaches a very high recall at the cost of precision whereas Noisemes SAD has the highest precision but the worst recall. Overall the F-scores of TO-Combo-SAD and OpenSMILE are close to each other, and much better than that of the Noisemes SAD, reflecting the pattern of the main WCE results. SAD F-scores did not correlate with the average SNR of each corpus ($p > 0.05$, Pearson correlation).

To further quantify how the SAD performance metrics affect WCE performance, Fig. 5 shows the WCE performance across all the six corpora as a function of the SAD precision, recall and F-score. As can be observed, increasing recall leads to lower error ($\rho = -0.45$, $p < 0.001$; rank correlation), but there is even a stronger effect of higher F-score leading to better performance ($\rho = -0.65$, $p < 0.001$). Together with Table 3, these results support the idea that SAD performance on the given data is related to the corresponding WCE accuracy.

It may seem surprising that the precision of SAD does not seem to correlate with WCE performance, and even recall explains only a limited proportion of the variance in the data. However, it is important to remember that the system is adapted to the given language in such a manner that any systematic under- or overestimation on the adaptation data is corrected with the α_L-parameter. Therefore the WCE error should converge to zero given enough data, as long as the test data have the same properties as the adaptation data. This also applies to the SAD: As long as performance of the SAD is similar in adaptation and testing, the error should diminish over time due to this correction mechanism. In contrast, if the recall or precision suddenly change due to adaptation, be that change an improvement or decline, it could be harmful to WCE performance. This is because the later stages of the system have no way of knowing if the distributional characteristics of the input coming have changed radically. Alternatively, sudden improvements in SAD performance could still boost the overall WCE accuracy: as the quality and/or quantity of target speech data captured by the SAD improves, the corresponding accuracy improvements in the uncorrected word count estimates may outweigh the problems caused by the changing distributional properties of the data.

In order to see whether changes in SAD performance (positive or negative) impact WCE errors, correlations between *adaptation-normalized SAD performance* and *corpus normalized WCE performance* were measured at the level of individual test participants, defined as follows: For TO-Combo-SAD and OpenSMILE SAD and each corpus, participant-specific SAD performance numbers (precision, recall, and F-score) were z-scored using the mean and variance of the adaptation data (i.e., excluding the subject itself from the dataset), and then transformed to the absolute value of the z-scores. As a result, 0 stands for test-case SAD performance typical to the adaptation data and positive values for increasingly different performance from the adaptation data. Participant-level relative WCE errors (Eq. (3)) were also z-scored in an analogous manner across all the test folds to quantify the error on a given participant compared to the performance on other subjects in the pool (negative value for below-average and positive value for above-average WCE error). The normalized errors were averaged across all the six syllabifiers, and the data

[8] Note that this is different from evaluating against all annotated speech segments (which further includes speech that is not comprehensible, speech by other children, and potential speech by the child with the recorder) or all annotated vocalizations (which include non-linguistic vocalizations).

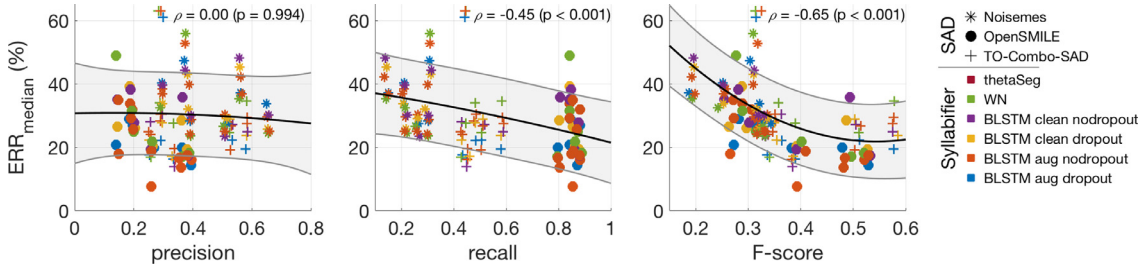

Fig. 5. WCE error as a function of SAD precision (left), recall (middle) and F-score (right). Different symbols indicate different SADs and colors indicate different syllabification algorithms. The shown 2nd order polynomial fit and the reported rank correlation r are calculated across all data points. A small amount of x-axis jitter has been added to the data to improve visual clarity.

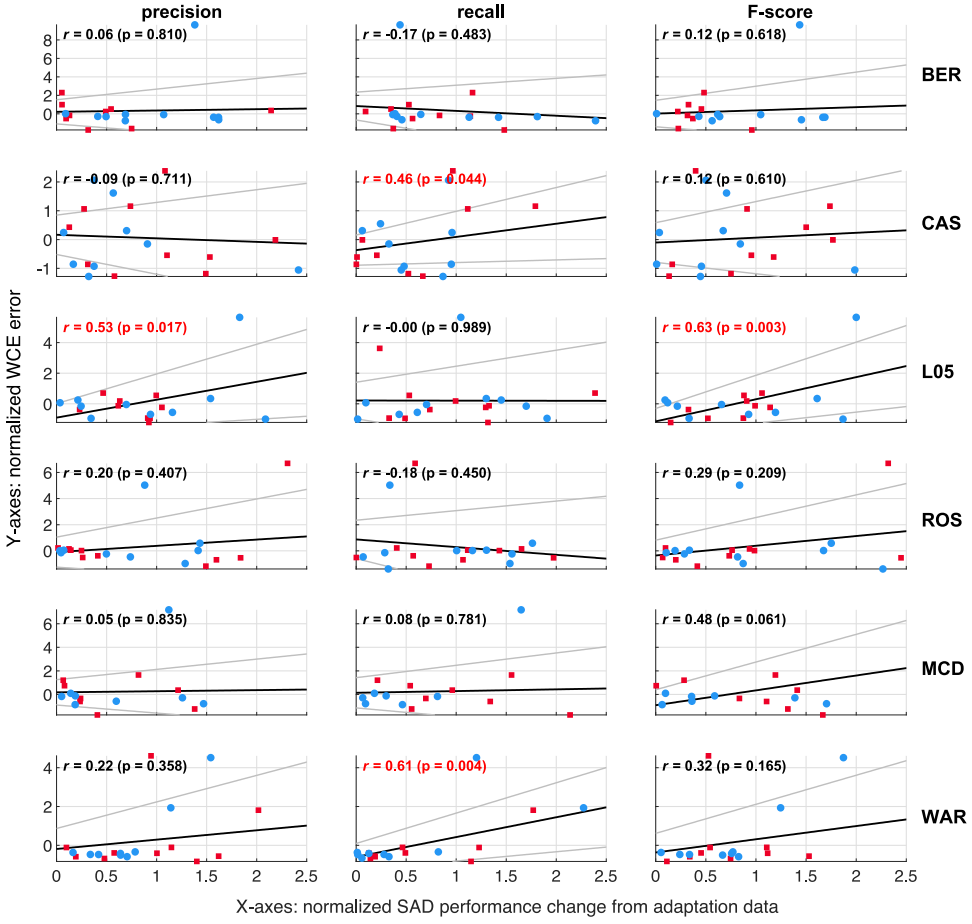

Fig. 6. Correlation between normalized WCE-performance (y-axis) and normalized absolute change in the SAD performance (x-axis) from adaptation data to test data. Red squares correspond to TO-Combo-SAD and blue circles to OpenSMILE. Significant correlations ($p < 0.05$) for pooled data from both SADs are highlighted in red. Each row corresponds to a different corpus, as labeled on the right.

from TO-Combo-SAD and OpenSMILE SAD were pooled before correlation calculation to focus on the shared effects of SAD behavior. Results of this analysis can be seen in Fig. 6.

The analysis with normalized data reveals that the changes in SAD performance between adaptation and testing do explain some of the WCE errors, but that the link between changes in SAD performance and WCE performance is not as straightforward as hypothesized above. Changes in recall correlate with WCE errors in CAS ($r = 0.46$, $p = 0.044$) and WAR ($r = 0.61$, $p = 0.004$), and now also precision change is correlated with WCE error in L05 ($r = 0.53$, $p = 0.017$). In addition, larger changes in F-score result in larger errors in L05 ($r = 0.63$, $p = 0.003$). However, precision, recall, and F-score changes have no effect in 14 out of the 18 cases investigated, and even the observed effects on L05 and CAS are relatively weak. In fact, only one the effects (WAR) would persist if a strict Bonferroni correction for multiple comparisons was carried out. This is despite nearly all corpora having several participants who have substantially different SAD behavior in testing than what has been observed during adaptation.

To see if the direction of SAD performance change is more informative of the WCE error, Fig. 7 shows the same analysis as above, but now using z-score normalized SAD scores but without taking the absolute value. In this case, negative SAD performance score means below- adaptation-average performance on the given test subject and positive SAD performance scores the opposite. Changes in precision have a clearer effect now: on L05, ROS, and WAR, participants with the largest errors also have notably worse precision than the rest of the participants, correlation reaching as high as $r = 0.74$ ($p < 0.001$) on L05. Conversely speaking, improvements in precision are associated with a decreasing error. A similar trend is also observed for the other three corpora, even though the effects are not significant. As for recall, there is

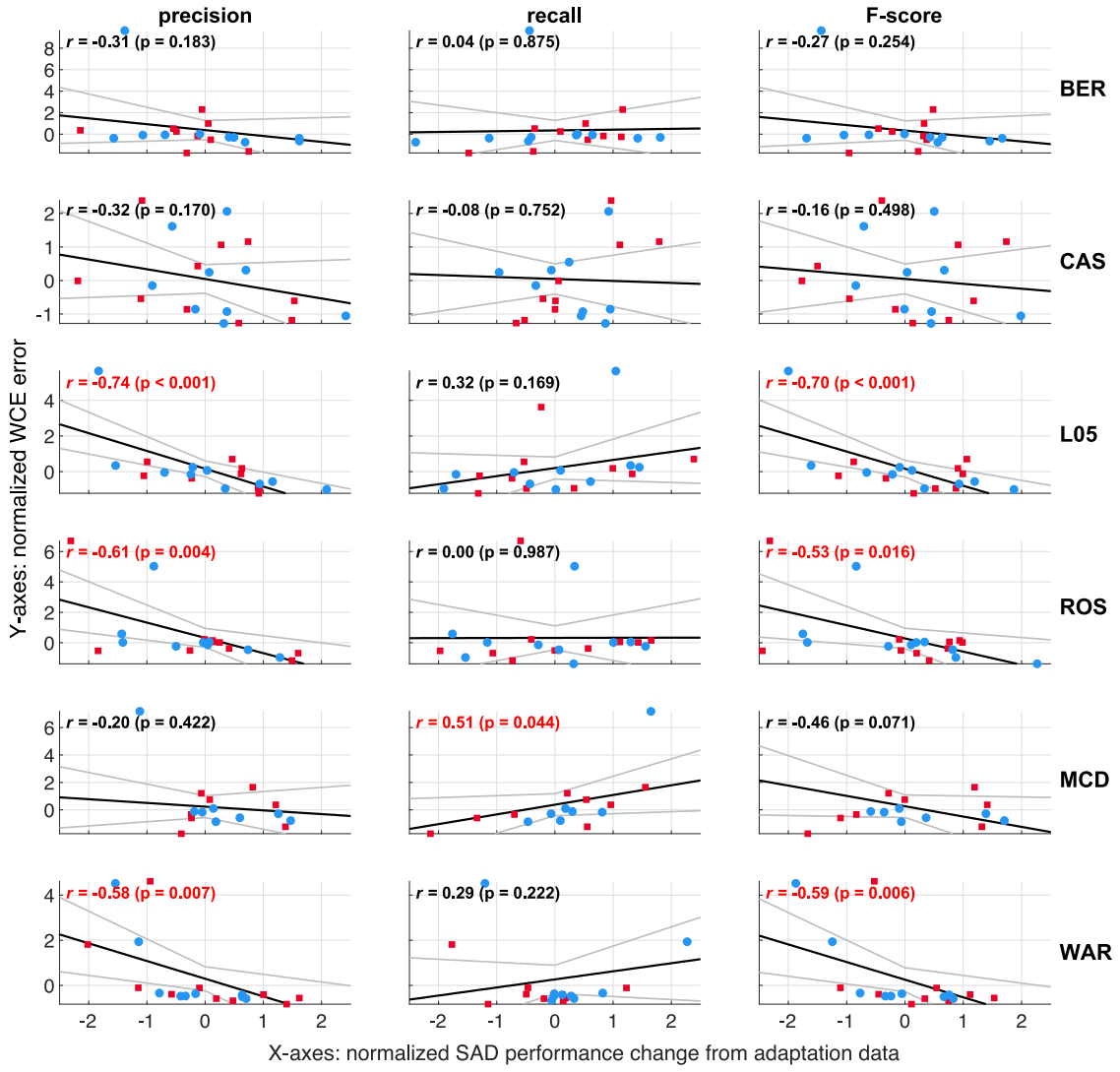

Fig. 7. Correlation between z-score normalized WCE-performance (*y*-axis) and *z*-score normalized change in the SAD performance (*x*-axis) from adaptation data to test data. Red squares correspond to TO-Combo-SAD and blue circles to OpenSMILE. Significant correlations ($p < 0.05$) or pooled data from both SADs are highlighted in red. Each row corresponds to a different corpus, as labeled on the right.

no longer an effect on WAR, but the two participants with the worst performance have either substantially lower or higher recall than majority of the population. For an unknown reason, MCD shows a pattern where worse recall is associated with a smaller WCE error. Overall, there is no clear pattern where decreases or increases in recall would map to systematic changes in WCE performance. Relative improvements in F-score are generally associated with better WCE performance, reaching significance on L05, ROS and WAR, but these seem to be largely driven by the improvements in precision.

Taken together, the results in Figs. 5–7 reveal that the SAD and WCE performance are partially connected. However, the connection is more complicated than simply stating that changing SAD performance would always map to worse WCE performance, or that any improvements in SAD performance would always lead to better WCE. What we can say is that (1) high and consistent SAD performance is naturally desired (see also top panel in Fig. 4 for performance if the SADs were ideal), (2) sometimes overall changes in recall from adaptation to testing are associated with larger estimation errors, and (3) improving precision from adaptation to test appears to be connected to improved performance. Still, it is important to remember that in all these cases, SAD and WCE performance numbers and their changes are ultimately reflecting some kind of qualitative properties of the audio signals themselves. A more detailed understanding of the sources of error would require a better understanding of what is actually happening in the audio data, but this type of analysis is beyond the scope of the present work.

4.4. Parameter variation across syllabifiers and languages

As a final step, we investigated how the adapted parameters θ_L and $\boldsymbol{\beta}_L$ vary across the tested syllabifiers and languages when using TO-Combo-SAD or OpenSMILE SAD. Full data on the analysis is shown in Appendix A, and the main findings can be summarized as follows:

(1) With TO-Combo-SAD, parameter β_1 controlling the relationship between estimated syllable counts and word counts performs as would be expected, having a similar value for the three well-performing English corpora (BER, MCD, L05). β_1 is lower for Tseltal and Spanish due to the higher number of syllables per word in these languages. For OpenSMILE SAD, there is a more complicated corpus-dependent pattern for the use of syllable counts and other alternative features in the word count predictions.

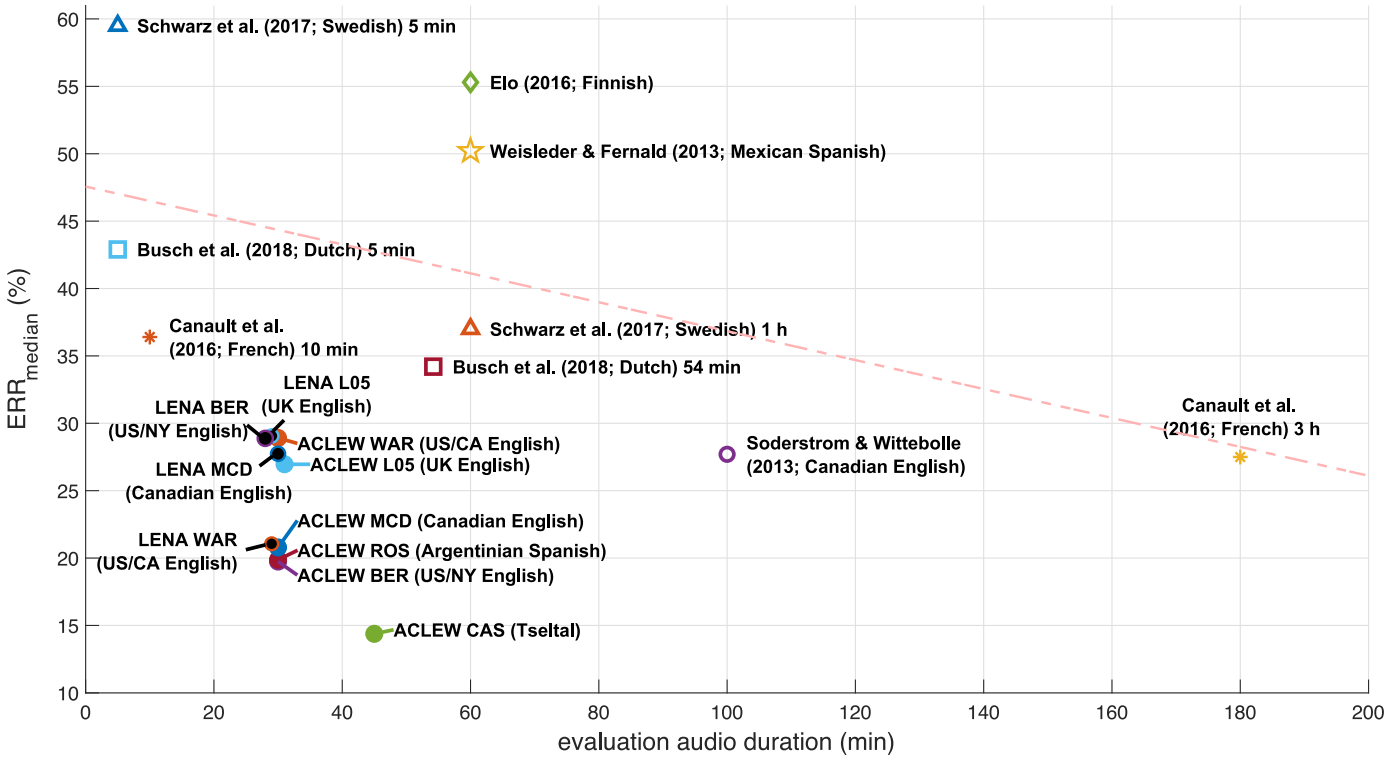

Fig. 8. Performance of the current WCE system ("ACLEW"; solid colored dots) plotted together with LENA performance on the same data (black dots) and in a number of earlier studies (other colored markers) using the ERR_{median} (%) as the performance metric (see also Table 1). The dashed red line shows a line fit to the data on LENA performance on non-English data. A slight amount of jitter is added to *x*-axis values to improve clarity.

(2) Syllable detection thresholds θ_L of the BLSTM-based syllabifier variants are similar across all corpora (approx. 0.6; remember that these values are nucleus likelihoods in range 0–1), suggesting that a fixed threshold could be used across languages. Variation of the optimal thresholds in WN and thetaSeg have much larger relative changes compared to the mean optimal value, corresponding to large qualitative changes between highly sensitive ($\theta_L \approx 0$) and much more conservative ($\theta_L > 0.1$) syllabification strategies (note that the absolute values are not directly comparable with the BLSTMs).

(3) In BLSTMs, speech duration is not used as a positive evidence for words, but is replaced by syllable count, mean sonority, and sometimes sonority SD. WN uses more varying weighing of the features depending on the corpus, sometimes using duration as a major predictor with more conservative syllabification strategy and vice versa.

(4) Parameter variation across training folds is typically limited within a corpus, as can be expected because roughly 90% of the data are the same in each training fold.

5. Conclusions and future work

The aim of this study was to describe a basic framework for automatic word count estimation from daylong audio recordings of sound environments of language-learning infants, and to test its applicability to multiple languages and language environments. We also compared a number of speech activity detectors and automatic syllabifiers as potential modules in the pipeline and studied the applicability of supervised neural network training for language-independent syllabification of speech, as evaluated in terms of overall WCE system performance.

One of the key aims was to have a system that performs similarly in high- and low-resource languages, as the existing commercially available LENA system is expected to perform better on English than other data. To place the present work in a context, Fig. 8 shows the current WCE system performance (with OpenSMILE SAD and augmented-training BLSTM syllabifier) together with LENA performance metrics on English from the same study. In addition, LENA performance on a variety of other languages from a number of earlier publications is shown. With the exception of the WAR corpus, the present system achieves lower relative word count error rates on all tested language samples compared to what LENA has achieved in the previous studies. It also outperforms LENA on three of the four varieties of English tested in the present experiments. Importantly, the performance on English, Argentinian Spanish, and Tseltal is very similar despite the wide variation in language and recorder types. In contrast, LENA accuracy has historically varied enormously depending on the language being recorded, with non-English data having substantially worse word count accuracies than English data (Fig. 8). This demonstrates that the basic approach consisting generic out-of-the-box SAD, language-independent syllabification, and domain-specific adaptation of a small number of parameters works both in principle and in practice. The current performance is far from ideal, but is still better than the static English-based acoustic model and mapping to word counts used in LENA. Unfortunately, our non-English corpora were not able to be processed with the LENA system, and therefore direct comparisons on exactly the same data is not possible.[9] It should also be noted that LENA word count estimates can also be linearly scaled to obtain more accurate word count estimates on novel languages and dialects, given that a suitable conversion coefficient is known or estimated from annotated data. In our experiments with the different varieties of English, linear scaling of LENA outputs based on the leave-one-subject-out adaptation protocol did not provide system-

[9] As mentioned in the introduction, LENA software processing is restricted to LENA-recorder outputs.

atic performance gains beyond the default LENA output (Section 4.1). However, LENA word count scaling is highly recommended for other languages if one wishes to use LENA to measure not only relative but also absolute word counts in an accurate manner.

From a technical point of view, the study demonstrates the applicability of supervised training of a neural network-based syllabifier. When multiple different training languages are used at the same time, such a syllabifier is also capable of reaching consistent behavior in languages not included in the training data. Furthermore, the BLSTM syllabifier outperformed two previously used syllabification methods, including the WN algorithm used in the WCE system of Ziaei et al. (2016), which had outperformed multiple alternative syllabifiers in the experiments of Ziaei et al. In addition, training data augmentation using a variety of realistic additive noise types and channel variability (based on random FIR-filters) was found to consistently improve syllabifier performance in the WCE task where signal conditions are extremely difficult compared to any typical speech processing problem. This suggests that neural network-based supervised syllabifiers could also work well in other tasks requiring syllable detection from speech. However, direct evaluation of syllabification accuracy instead of WCE performance was beyond the scope of the present study, and should be carried out separately.

5.1. Limitations and future work

The pattern of results shows that basic idea of adapting the system to a new language or dialect by using 30–60 min of annotated adult speech works in principle, and that the WCE performance does not critically depend on the language in question. Instead, the main performance problems, especially transparent on the WAR data, seem to be associated with at least two central factors: (1) the current lack of a reliable component for separating different sources and styles of vocal activity from each other, and (2) limited SAD performance on the daylong data.

In the present WCE system, all speech passing through a generic SAD is treated as equal, whereas for child language research it would be important to distinguish adults, siblings, the key child, and, e.g., sources of electronic speech (TV, radio) from each other since prior work shows that child-directed speech, particularly from adults, is predictive of children's later linguistic development (e.g., Shneidman and Goldin-Meadow, 2012). In addition, some content such as singing or non-linguistic communicative vocalizations (e.g., laughter) can be categorized as speech, but its acoustic features do not have the same relationship to spoken word counts that normal speech does. To allow direct comparison with LENA, we chose to only evaluate our system against transcribed speech from adult talkers. However, now that the basic concept and its functionality have been validated, the next efforts should be directed toward the development and testing of a robust speaker diarization module required for speaker attribution. Although the current ACLEW virtual machine published in Le Franc et al. (2018) already contains one such a tool, DiarTK (Vijayasenan and Valente, 2012), its performance was found to be lacking on child daylong data (see also DiHARD diarization challenge[10] where DiarTK scored at the bottom among all the submissions; see also Le Franc et al., 2018). In order to maintain focus on SAD and syllabifier comparisons, no separate experiments with diarization tools were included in the present report. More work is needed to identify the best ways to tackle the problem of who is speaking (and whether it is an electronic device or a live person), and preferably also what the style of speech is (infant-directed, adult-directed, singing, shouting, etc.).

It is also obvious that the performance of all compared SADs is far from ideal on the present type of daylong data, as WCE from SAD outputs falls far behind the performance of a system using oracle utterance boundaries. This problem could be approached in several ways in the future. One option is to start using a SAD that can also be adapted to the target domain, or at least to re-train the current SADs on a large amount of child daylong recordings instead of using the original models provided by the algorithm authors. An alternative solution would be to integrate SAD with the speaker diarization or syllabification algorithms, and seek ways to efficiently train or adapt this unified model to the daylong data. Speech enhancement and/or statistical normalization as a front-end for SAD should be also investigated, as the strategy used so far was to apply enhancement only to SAD output segments in order to save computational costs. In general, more intelligent, robust, and adaptable systems towards the highly variable signal conditions and signal statistics encountered in unconstrained daylong recordings are required.

The final important factor to mention is that the gold standard word counts derived from orthographic annotations are not always unanimous due to several factors, and this problem applies to WCE evaluation both in the current system and in LENA. In the current corpora, all speech that the annotators could transcribe based on repeated listening had been transcribed, while the unclear vocalizations are simply marked as "cannot transcribe". In practice, there is a continuum from clear near-field speech to hardly audible noisy content that is only partially comprehensible. Since the current datasets do not provide any information on the clarity of the input for the transcribed words, all transcribed tokens from this continuum are treated as equally relevant targets for the WCE. Another potential source of uncertainty comes from the potential differences in how faithfully spoken language maps into orthographic transcripts, and especially how well whitespaces in the orthographic transcripts can be used to define word boundaries in the running speech in different languages compared. Importantly, however, there is no obvious reason why uncertainties in the gold standard word counts would specifically favor any of the compared system configurations. Instead, it is simply important to be aware that the performance figures for word count accuracies hide a number of factors that may artificially inflate or deflate the error rates, ultimately depending on how the actual target word counts are transcribed and defined for the evaluations.

In sum, this study is the first to test publicly available speech tools for word count estimation on daylong child recordings in different languages in comparable settings, and, to the best of our knowledge, the first to demonstrate the applicability of language-independent supervised syllabifiers of speech. More work is still needed to come up with a comprehensive set of open-source tools for analyzing linguistic content in daylong real-world recordings. The present relatively straightforward system for word count estimation is only the first step in that direction.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded as a part of *Analyzing Child Language Experiences around the World* (ACLEW) collaborative project funded by the Trans-Atlantic Platform for Social Sciences and Humanities "Digging into Data" challenge, including a local Academy of Finland grant (312105) to OR, ANR-16-DATA-0004 ACLEW to AC, NEH HJ-253479-17 to EB, HJ-253479 to CR, and funding from the Social Sciences and Humanities Research Council of Canada to MS (869-2016-0003). MC was funded by an NWO Veni Innovational Research grant (275-89-033). In addition, OR was funded by an Academy of Finland grant no. 314602, MS by a Social Sciences and Humanities Research Council of Canada Insight Grant (435-2015-0628), and AC by Agence Nationale de la Recherche (ANR-14-CE30-0003 MechELex, ANR-17-EURE-0017) and the J. S. McDonnell Foundation Understanding Human Cognition Scholar Award. EB was funded by NIH DP5-OD019812, and CR by CONICET grants PIP 80/201 and PICT 3327/2014.

[10] https://coml.lscp.ens.fr/dihard/index.html.

The authors would like to thank Tobias Busch, Adriana Weisleder, Iris-Corinna Schwarz, Ellen Marklund, and Melanie Canault and her colleagues for providing their data on LENA reliability on the various languages, and John Hansen for kindly providing the TO-Combo-SAD algorithm to be used in the project, as well as Anne Warlaumont and the LuCiD Language0–5 team for generously providing access to their corpora. Thanks are also due to the many research assistants who contributed to the annotation and transcription of the corpora, and to the many research participant families who generously provided audio access to their everyday intimate lives for research purposes.

An early version for parts of this work was presented at Interspeech-2018 conference in September 2018 at Hyderabad, India.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.specom.2019.08.005.

Appendix A: Parameter variation across corpora and subjects

Fig. A1 shows the linear mapping parameter values across different syllabifiers and corpora when using TO-Combo-SAD. Fig. A2 shows the same parameters for OpenSMILE SAD.

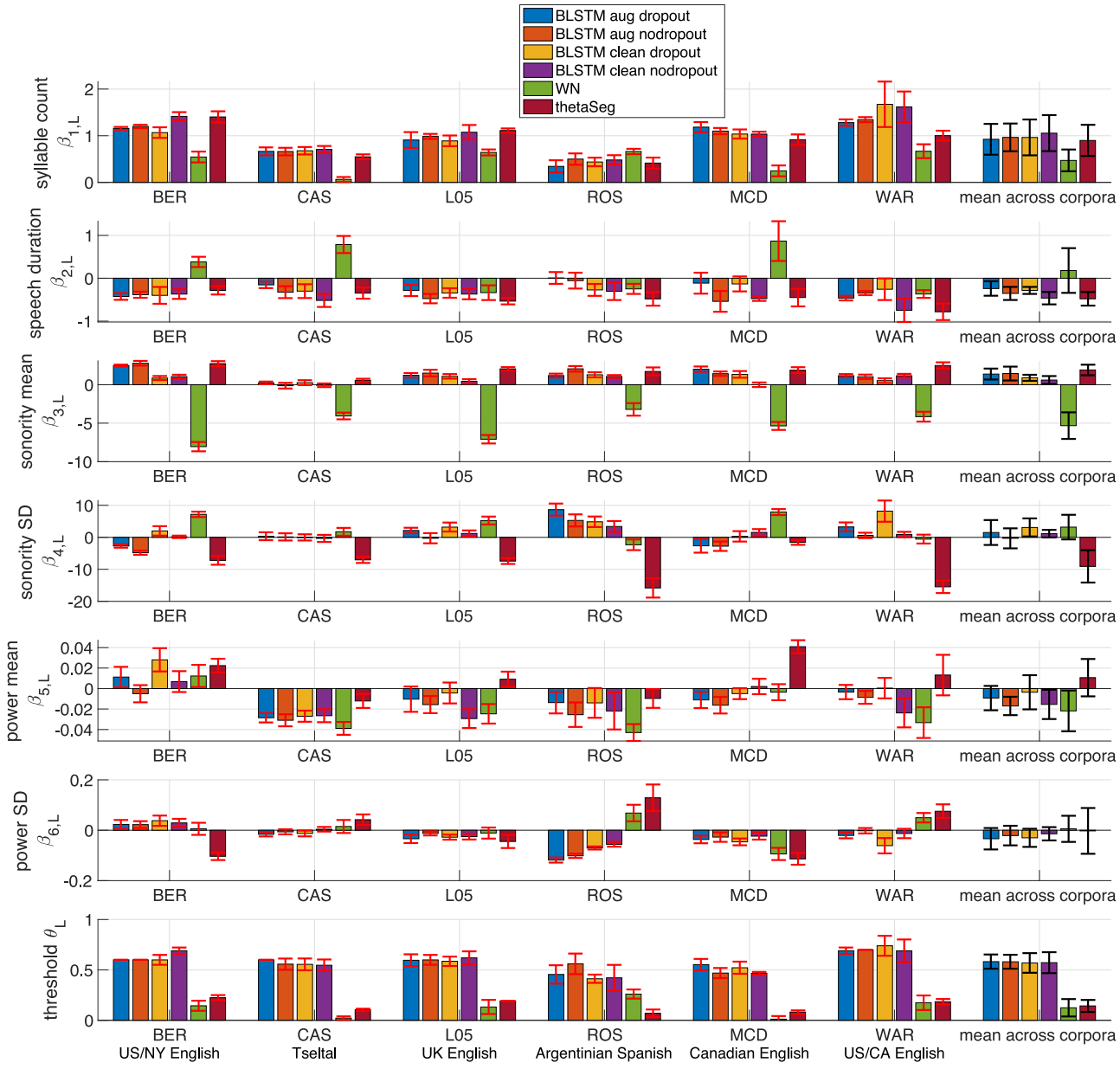

Fig. A1. Learned mapping parameters θ_L and $\boldsymbol{\beta}_L$ of different syllabifiers on the different corpora (with TO-Combo-SAD). Different color bars stand for different syllabifiers (see the legend) and red error bars denote parameter standard deviation across all training folds on the given corpus. Mean parameter values across all the corpora are shown, for which standard deviation of the parameter across the corpora is shown with black error bars. Top panel: number of detected syllables per word. Second panel: duration of speech (in seconds) per word. Middle panel: mean of the sonority envelope. Fourth panel: SD of sonority envelope. Fifth panel: mean of signal power. Sixth panel: SD of signal power. Bottom panel: syllable detection threshold.

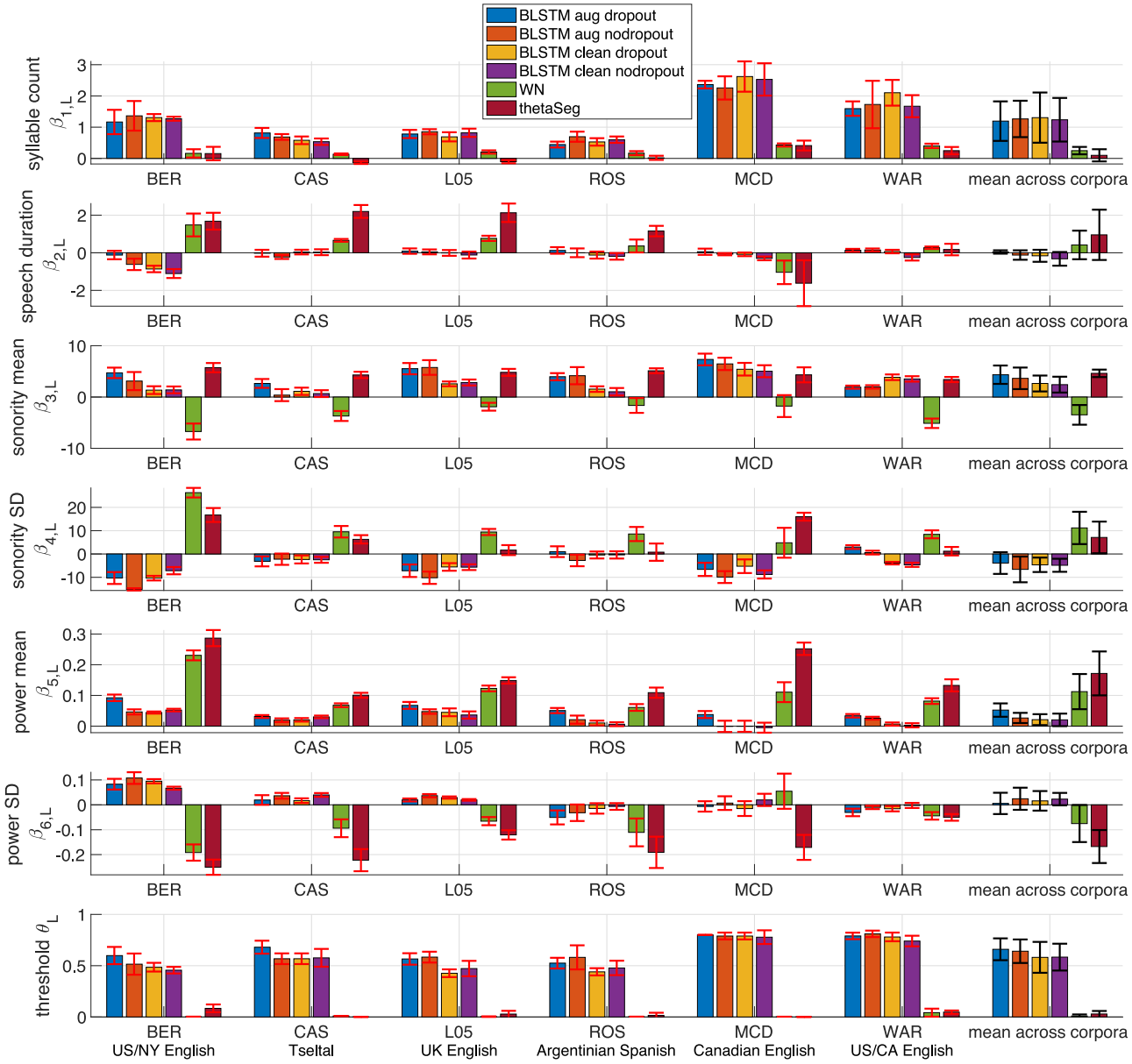

Fig. A2. Learned mapping parameters θ_L and $\boldsymbol{\beta}_L$ of different syllabifiers on the different corpora (with OpenSMILE SAD).

References

Avanzi, M., Simon, A., Goldman, P.-J., Auchlin, A., 2010. C-PROM: an annotated corpus for French prominence study. In: Proceedings of the Speech Prosody, Chicago, IL, pp. 10–14 May pp..

Bergelson, E. (2016). Bergelson seedlings homebank corpus. doi:10.21415/T5PK6D.

Bergelson, E., Warlaumont, A., Cristia, A., Casillas, M., Rosemberg, C., Soderstrom, M., Rowland, C., Durrant, S., Bunce, J., 2017a. Starter-ACLEW. Databrary Retrieved November 9, 2018 from doi:10.17910/B7.390.

Bergelson, E., Cristia, A., Soderstrom, M., Warlaumont, A., Rosemberg, C., Casillas, M., Rowland, C., Durrant, S., Bunce, J., 2017b. ACLEW Project. Retrieved November 1, 2018 from https://nyu.databrary.org/volume/389 .

Bergelson, E., Casillas, M., Soderstrom, M., Seidl, A., Warlaumont, A.S., Amatuni, A., 2018a. What do north American babies hear? A large-scale cross-corpus analysis. Dev. Sci. 22 (1), e12724 electronic pre-print.

Bergelson, E., Amatuni, A., Dailey, S., Koorathota, S., Tor, S., 2018b. Day by day, hour by hour: naturalistic language input to infants. Dev. Sci. 22, e12715.

Berouti, M., Schwartz, R., Makhoul, J., 1979. Enhancement of speech corrupted by acoustic noise. In: Proceedings of the ICASSP-79, Washington, DC, pp. 208–211 April 2–4.

Brent, M.R., Siskind, J.M., 2001. The role of exposure to isolated words in early vocabulary development. Cognition 81, 31–44.

Canault, M., Le Normand, M-T., Foudil, S., Loundon, N., Thai-Van, H., 2016. Reliability of the Language Environment Analysis system (LENA™) in European French. Behav. Res. Methods 48 (3), 1109–1124.

Casillas, M., Brown, P., & Levinson, S.C. (2017). Casillas homebank corpus. https://homebank.talkbank.org/access/Secure/Casillas.html

Casillas, M., Bergelson, E., Warlaumont, A., Cristia, A., Soderstrom, M., VanDam, M., Sloetjes, H., 2017a. A new workflow for semi-automatized annotations: tests with long–form naturalistic recordings in children's language environments. In: Proceedings of the Interspeech-2017, Stockholm, Sweden, pp. 2098–2102 August 20–24.

Casillas, M., Bunce, J., Soderstrom, M., Rosemberg, C., Migdalek, M., Alam, F., Stein, A., & Garrison, H. (2017b). Introduction: the ACLEW DAS template. Online material available at https://osf.io/aknjv/.

Clements, G.N., 1990. The role of the sonority cycle in core syllabification. In: Kingston, J., Beckman, M.E. (Eds.), Papers in Laboratory Phonology 1: Between the Grammar and Physics of Speech. Cambridge University Press, Cambridge, pp. 283–333.

Cristia, A., Dupoux, E., Gurven, M., Stieglitz, J., 2017. Child-Directed speech is infrequent in a forager-farmer population: a time allocation study. Child Dev. online pre-print doi:10.1111/cdev.12974.

de Saussure, F., 1916. Cours De Linguistique Générale. Payot, Paris.

Elo, H., 2016. Acquiring language as a twin: twin children´s early health, social environment and emerging language skills. Acta Universitatis Tamperensis 2240. Tampere University Press, Tampere. Doctoral thesis.

Eyben, F., Weninger, F., Squartini, S., Schuller, B., 2013a. Real-life voice activity detection with LSTM recurrent neural networks and application to hollywood movies. In: Proc. ICASSP-2013, Vancover, Canada, pp. 483–487 May 26–31.

Eyben, F., Weninger, F., Gross, F., Schuller, B., 2013b. Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In: Proceedings of the ACM Multimedia (MM), Barcelona, Spain, pp. 835–838 October 21–25.

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

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., 1990. The darpa timit acoustic–phonetic continuous speech corpus. national institute of standards and technology speech. Disc 1–1.1, NTIS Order No. PB91–505065.

Gilkerson, J., Richards, J., 2009. The LENA Natural Language Study. LENA Foundation, pp. 1–26 Technical Reports (September 2008).

Gilkerson, J., Zhang, Y., Xu, D., Richards, J., Xu, X., Jiang, F., Harnsberger, J., Topping, K., 2015. Evaluating language environment analysis system performance for chinese: a pilot study in Shanghai. J. Speech Lang. Hear. Res. 58, 445–452.

Godfrey, J.J., Holliman, E.C., McDaniel, J., 1992. SWITCHBOARD: telephone speech corpus for research and development. In: Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing, San Francisco, CA, pp. 517–520.

Hart, B., Risley, T.R., 1995. Meaningful Differences in the Everyday Experience of Young American Children. Paul H Brookes Publishing, Baltimore, MD.

Hoff, E., 2006. How social contexts support and shape language development. Dev. Rev. 26 (1), 55–88.

Henrich, J., Heine, S.J., Norenzayan, A., 2010. The weirdest people in the world? Behav. Brain Sci. 33 (2–3), 61–83.

Hermansky, H., Morgan, N., 1994. RASTA processing of speech. IEEE Trans. Speech Audio Process. 2 (4), 578–589.

Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J., Hedges, L.V., 2010. Sources of variability in children's language growth. Cogn. Psychol. 61 (4), 343–365.

Landsiedel, C., Edlund, J., Eyben, F., Neiberg, D., Schuller, B., 2011. Syllabification of conversational speech using bidirectional long-short-term memory neural networks. In: Proceedings of the ICASSP-2011, Prague, Czech Republic, pp. 5256–5259 May 22–27.

Le Franc, A., Riebling, E., Karadayi, J., Wang, Y., Scaff, C., Metze, F., Cristia, A., 2018. The aclew divime: an easy-to-use diarization tool. In: Proceedings of the Interspeech-2018, Hyderabad, India, pp. 1383–1387 September 2–6.

Lippus, P., Tuisk, T., Salveste, N., Teras, P., 2013. Phonetic Corpus of Estonian Spontaneous Speech. Institute of Estonian and General Linguistics, University of Tartu doi:10.15155/TY.000D.

Lieven, E., 1994. Crosslinguistic and crosscultural aspects of language addressed to children. In: Gallaway, C., Richards, B.J. (Eds.), Input and Interaction in Language Acquisition. Cambridge University Press, Cambridge, pp. 56–73.

Martin, R., 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. 9 (5), 504–512.

McDivitt, K., & Soderstrom, M. (2016). McDivitt homebank corpus. doi:10.21415/T5KK6G.

Mermelstein, P., 1975. Automatic segmentation of speech into syllabic units. J. Acoust. Soc. Am. 58, 880–883 1975.

Metze, F., Fosler-Lussier, E., Bates, R., 2013. The speech recognition virtual kitchen. In: Proceedings of the Interspeech-2013, Lyon, France, pp. 1858–1860 August 25–29.

Morgan, N., Fosler-Lussier, E., 1998. Combining multiple estimators of speaking rate. In: Proceedings of the ICASSP-98, Seattle, WA, pp. 729–732 May 12–15.

Obin, N., Lamare, F., Roebel, A., 2013. Syll-O-Matic: an adaptive time-frequency representation for the automatic segmentation of speech into syllables. In: Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP-2013), Vancouver, BC, pp. 6699–6703.

Parker, S.G., 2002. Quantifying the Sonority Hierarchy. Graduate School of the University of Massachusetts, Amherst, MA.

Pitt, M.A., Johnson, K., Hume, E., Kiesling, S., Raymond, W., 2005. The buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. Speech Commun. 45, 89–95.

Plummer, A., Riebling, E., Kumar, A., Metze, F., Fosler-Lussier, E., Bates, R., 2014. The speech recognition virtual kitchen: launch party. In: Proceedings of the Interspeech-2014, Singapore, pp. 2140–2141 September 14–18.

Price, P.J., 1980. Sonority and syllabicity: acoustic correlates of perception. Phonetica 37, 327–343.

Ramírez-Esparza, N., García-Sierra, A., Kuhl, P.K., 2014. Look who's talking: speech style and social context in language input to infants are linked to concurrent and future speech development. Dev. Sci. 17 (6), 880–891.

Rosemberg, C.R., Alam, F., Stein, A., Migdalek, M., Menti, A., & Ojea, G. (2015). Los entornos lingüísticos de niñas y niños pequeños argentinos / language environments of young argentinean children. CONICET (DOI in progress).

Rowe, M.L., 2012. A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. Child Dev. 83 (5), 1762–1774.

Rowland, C.F., Bidgood, A., Durrant, S., Peter, M., Pine, J.M., 2018. The Language 0-5 Project. University of Liverpool Unpublished manuscript Available from https://osf.io/kau5f/ doi:10.17605/OSF.IO/KAU5F.

Rytting, A., Brew, C., Fosler-Lussier, E., 2010. Segmenting words from natural speech: subsegmental variation in segmental cues. J. Child Lang. 37, 513–543.

Räsänen, O., Doyle, G., Frank, M.C., 2018. Pre-linguistic segmentation of speech into syllable-like units. Cognition 171, 130–150.

Sadjadi, S., Hansen, J., 2013. Unsupervised speech activity detection using voicing measures and perceptual spectral flux. IEEE Signal Process. Lett. 20 (3), 197–200.

Schwarz, I-C., Botros, N., Lord, A., Marcusson, A., Tidelius, H., Marklund, E., 2017. The LENA^{TM} system applied to Swedish: reliability of the adult word count estimate. In: Proceedings of the Interspeech-2017. Stockholm, Sweden, pp. 2088–2091 August 20–24.

Shneidman, L., Goldin-Meadow, S., 2012. Language input and acquisition in a Mayan village: how important is directed speech? Dev. Sci. 15 (5), 659–673.

Soderstrom, M., Wittebolle, K., 2013. When do caregivers talk? The influences of activity and time of day on caregiver speech and child vocalizations in two childcare environments. PLoS One 8 (11), e80646.

Strassel, S., Morris, A., Fiscus, J.G., Caruso, C., Lee, H., Over, P.D., Fiumara, J., Shaw, B.L., Antonishek, B., Michel, M., 2012. Creating HAVIC: heterogeneous audio visual internet collection. In: Proceedings of the LREC-2012., Istanbul, Turkey, pp. 2573–2577 May 21–27.

Tamis-LeMonda, C., Kuchirko, Y., Luo, R., Escobar, K., Bornstein, M., 2017. Power in methods: language to infants in structured and naturalistic contexts. Dev. Sci. 20 (6), e12456.

VanDam, M., Warlaumont, A.S., Bergelson, E., Cristia, A., Soderstrom, M., Palma, P.D., MacWhinney, B., 2016. HomeBank: an online repository of daylong child-centered audio recordings. Semin. Speech Lang. 37 (2), 128–142. doi:10.1055/s-0036-1580745.

Vijayasenan, D., Valente, F., 2012. DiarTk: an open source toolkit for research in multistream speaker diarization and its application to meetings recordings. In: Proceedings of the Interspeech-2012. Portland, OR September 9–13.

Villing, R., Timoney, J., Ward, T., Costello, J., 2004. Automatic blind syllable segmentation for continuous speech. In: Proceedings of the Irish Signals and Systems Conference (ISSC 2004). Belfast, Northern Ireland.

Wang, D., Narayanan, S., 2007. Robust speech rate estimation for spontaneous speech. IEEE Trans. Audio Speech Language Process. 15 (8), 2190–2201.

Wang, Y., Neves, L., Metze, F., 2016. Audio-based multimedia event detection using deep recurrent neural networks. In: Proceedings of the ICASSP-2016, Shanghai, China, pp. 2742–2746 March 20–25.

Warlaumont, A.S., Pretzer, G.M., Mendoza, S. & Walle, E.A. (2016). Warlaumont homebank corpus. doi:10.21415/T54S3C

Weisleder, A., Fernald, A., 2013. Talking to children matters: early language experience strengthens processing and builds vocabulary. Psychol. Sci. 24 (11), 2143–2152.

Whitney, W.D., 1874. Oriental and Linguistic Studies. Scribner, Armstrong & Co, New York Second Series.

Xu., D., Yapanel, U., Gray, S., Gilkerson, J., Richards, J., Hansen, J., 2008. Signal processing for young child speech language development. In: Proceedings of the 1st Workshop on Child Computer and Interaction (WOCCI-2008). Chania Crete, Greece October 23.

Yun, W., Yoon, K., Park, S., Lee, J., Cho, S., Kang, D., Byun, K., Hahn, H., Kim, J., 2015. The Korean corpus of spontaneous speech. Phon. Speech Sci. 7 (2), 103–109.

Ziaei, A., Sangwan, A., Kaushik, L., Hansen, J., 2015. Prof-life-log: analysis and classification of activities in daily audio streams. In: Proceedings of the ICASSP-2015, Brisbane, Australia, pp. 4719–4723 April 19–24.

Ziaei, A., Kaushik, L., Sangwan, A., Hansen, J., 2014. Speech activity detection for NASA apollo space missions: challenges and solutions. In: Proceedings of the Interspeech-2014, Singapore, pp. 1544–1548 Sept. 14–18.

Ziaei, A., Sangwan, A., Hansen, J., 2016. Effective word count estimation for long duration daily naturalistic audio recordings. Speech Commun. 84, 15–23.

Ziaei, A., Sangwan, A., Hansen, J.H.L., 2013. Prof-life-log: personal interaction analysis for naturalistic audio streams. In: Proceedings of the ICASSP-2013, Vancouver, Canada, pp. 7770–7774 May 16–23.