A Speech System for Estimating Daily Word Counts
Ali Ziaei, Abhijeet Sangwan, John H.L. Hansen
Center for Robust Speech Systems (CRSS),
Department of Electrical Engineering, University of Texas at Dallas, Richardson, Texas, U.S.A
{ali.ziaei,abhijeet.sangwan,john.hansen}@utdallas.edu

Abstract
The ability to count the number of words spoken by an individual over long durations is important to researchers investigating language development, healthcare, education, etc. In this study, we attempt to build a speech system that can compute daily word counts using data from the Prof-Life-Log corpus. The task is challenging as typical audio files from Prof-Life-Log tend to be 8-to-16 hours long, where audio is collected continuously using the LENA device. This device is worn by the primary speaker and all his daily interactions are collected in fine detail. The recordings contain a wide variety of noise types with varying SNR (signal-to-noise ratio) including large crowd, babble, and competing secondary speakers. In this study, we develop a word-count estimation (WCE) system based on syllable detection and we use the method proposed by Wang and Narayanan as the baseline system [1]. We propose many modifications to the original algorithm to improve its effectiveness in noise. Particularly, we incorporate speech activity detection and enhancement techniques to remove non-speech from analysis and improve signal quality for superior syllable detection, respectively. We also investigate features derived from syllable detection for better word count estimation. The proposed method show significant improvement over the baseline.

Index Terms: Word Count Estimation, Naturalistic Audio, Noise Robustness, Vowel Detection, Speaking Rate

1. Introduction

Prof-Life-Log is a unique speech corpus which contains continuous long duration audio recordings. In this collection, the primary speaker wears the audio recording unit throughout the workday. Data is captured continuously as the primary speaker navigates through the daily work schedule, and typical recordings are 8-to-16 hours long. In this manner, the captured data contains large amounts of information about the speaker and his environment. Therefore, the collection opens up the opportunity to ask some interesting questions related to speaker, speech, environment and language [2, 3, 4, 5, 6]. One such question is: “how many words did the primary speaker speak today?” If this study, we attempt to build a system that can answer this question.

The ability to count the number of words spoken by an individual over long durations of time is interesting and important to a number of fields. The importance of word count in language acquisition and development has been well researched. For example, in [7], the authors discuss three main findings, i.e., (i) the variation in children’s IQs and language abilities is relative to the amount parents speak to their children, (ii) children’s academic successes at ages nine and ten are attributable to the amount of talk they hear from birth to age three. and (iii) the parents of advanced children talk significantly more to their children than parents of children who are not as advanced. More recently, word count estimates are also being researched in child-Autism studies [8].

In general, we can adopt two broad approaches towards building word count estimation (WCE) systems. The first approach relies on performing automatic speech recognition followed by simple word counter. However, the ASR method must address a number of challenges in order to provide robust measurements across a wide variety of conditions. Naturalistic long duration audio recordings (such as Prof-Life-Log) contains a wide variety of noise conditions that directly impact speech recognition performance. Additionally, in practical settings, ASR must also contend with non-native speakers, foreign accents and bilingual/multilingual speakers. Finally, ASR systems are computationally demanding and may not be suitable for all use cases. An alternative to speech recognition is to use signal processing methods to discover low level structures in speech (such as vowels, phones or syllables). The problem of identifying low level structures in speech by detecting periodicity, harmonics and/or formants has been well investigated by researchers and a number of methods have been proposed. This rich repository of ideas forms a nice platform to build a viable word counting estimation system. In this study, we develop a WCE system based on syllable detection. Particularly, we used the method proposed by Wang and Narayanan as the baseline system, which in turn is a modification of *nrate* method (and has been used to measure speech rate) [1, 9].

We propose a number of additions to the baseline system in order to develop a WCE system that works for Prof-Life-Log data. Since the audio recordings are very long, we segment the file into approximately 1-minute continuous segments and run word count estimation independently for each segment before combining the estimates to obtain a single final number. In order to estimate word count, we construct a simple yet effective LMMSE (Linear Minimum Mean Square Error) estimator for word rate from syllable rate.

Our analysis of the LMMSE estimator reveals that the performance in very sensitive to noise. In Prof-Life-Log, we encounter a wide variety of environments such as in-vehicle, large-crowd, outdoors, office, etc., and this poses a significant noise robustness challenge. To address this problem, we employ Speech Activity Detection (SAD) to assist in removing non-speech and focus syllable detection on speech regions alone. Particularly, we employ the UTD Combo-SAD approach which has shown very good performance in the recent DARPA RATS (robust automatic transcription of speech) evaluations [10].

While SAD helps in removing non-speech regions from
processing, the accuracy of syllable detection on noisy speech remains poor. In order to mitigate this problem, we apply popular speech enhancement techniques to produce enhanced speech for syllable detection. Our experiments show that this approach is very effective in improving the accuracy of syllable detection which directly impacts WCE.

2. Prof-Life-Log

We use the LENA device to capture audio data in Prof-Life-Log. LENA is a light weight compact digital audio recorder that is capable of capturing data continuously for about 16 hours. One of the primary applications of LENA is to record the language environments of infants and young children, where the young subject wears the unit [11].

2.1. Data Selection

In Prof-Life-Log, the LENA device is worn by a subject (also called the primary speaker) for the entire workday, and the data is captured continuously. Currently, we have collected data for more than 50+ work days. Most of the collection is from typical work days at UTD, but some days have been collected while the subject was attending conferences, workshops, etc. For this study, we chose large portions from 2 full-day recording (approximately 6 hours each). One 6-hour segment was chosen from a typical workday at the University and the other 6-hour segment was chosen from a conference workday (giving us a total of over 12 hours of evaluation data). The typical workday recording contains speech in environments such as classroom, cafeteria, campus outdoors, lab, hallways etc. The conference day recording is much more noisier, and it contains large crowd noise, babble noise, and competing speakers (i.e., overlapped speech with primary speaker).

For system evaluation, this data was cut into approximately 1-minute segments and each cut was transcribed. Additionally, environment labels such as office, classroom, outdoors, cafeteria, etc. were applied for each 1-minute cut. Finally, primary and secondary speaker labels were also applied (all speakers other than the primary speaker were labelled as secondary speaker). It is noted that the objective of this study is to estimate the primary speaker’s word count estimates.

2.2. Data Analysis

From WCE perspective, Prof-Life-Log poses two significant challenges: (i) wide variety of noise types and time varying SNR (signal-to-noise ratio), and (ii) secondary speakers. Figure 1 shows the NIST-STNR and WADA-SNR histogram plots for Prof-Life-Log data [12, 13]. It is seen that the NIST-STNR and WADA-SNR distributions are bimodal and trimodal, respectively. The higher SNR values reflect data captured in cleaner environments (such as office) and the lower SNR values reflect noisier environments (such as cafeteria).

Secondary speakers constitute the second major challenge for WCE. They come in two flavors: (i) non-competing secondary speakers who speak when the primary speaker is not speaking (these are generally people who are talking to the primary speaker), and (ii) competing secondary speakers who speak when the primary speaker speaks (these are generally people who are having parallel conversations which is common in crowded environments). Relatively speaking, competing secondary speakers are a bigger problem. Fortunately, they are also rare when compared to non-competing secondary speakers. In order to provide a more objective understanding of this problem, we compute the proportion of primary and secondary speech in every cut. We call these numbers primary and secondary speaker duration ratios. If the primary speaker duration ratio is 1, then the primary speaker alone spoke in the entire cut. Similarly, if secondary speaker duration ratio is 1, then the secondary speaker alone spoke in the entire cut. Figure 2 shows the scatter plot of primary and secondary speaker duration ratios. The scatter plot illustrates that the data consists of cuts where either the primary or the secondary speaker dominates. Additionally, there are also cuts where the primary and secondary speech is balanced. Finally, there are cuts where neither dominates (i.e., cut contains mostly non-speech). It is also important to note that the intelligibility of secondary speech in Prof-Life-Log is also variable (and this can be attributed to distance from the microphone and/or environmental noise). It ranges from perfectly intelligible speech to a more babble-like quality.

In this study, we focus on the noise robustness problem alone. Therefore, we ignored cuts from Prof-Life-Log that contained secondary speakers, and only selected primary speaker cuts for the evaluation of the proposed method. In what follows, we describe our system and demonstrate how our method systematically reduces errors due to noisy environment.

3. Word Count Estimation (WCE) System

Our baseline system uses the syllable detection technique proposed by Wang and Narayanan [1]. This technique is based on the mrate method, and it relies on peak-picking to detect
syllables in the speech signal. The peak-picking process itself uses a smoothed spectrogram and pitch estimates. The basic algorithm outputs syllable counts which are easily converted into syllable rate. However, as we are interested in word counts, we employ a simple LMMSE (linear minimum mean square error) estimator for computing word rate (and therefore word count) from syllable rate. For our experiments, we learned the LMMSE parameters on a held out dataset derived from the Switchboard corpus. The combination of the syllable detector (A) and LMMSE estimator (B) constitutes our baseline WCE system.

Using experimental evaluation, we confirmed that the baseline system is extremely susceptible to noise, and the performance deteriorates rapidly for challenging noise types (e.g., large-crowd) and low SNR values. Systematic analysis of the data revealed that the syllable detector was frequently generating false-alarms for noise alone. This was especially true for noise which contains harmonics (such as machine noise), and speech-like noise (i.e., babble, large-crowd, etc.). A simple solution to reduce false-hits was to add a speech activity detection (SAD) system, and process speech-only regions for WCE. This approach is frequently used in ASR systems to reduce insertion rates. In this study, we employed the Combo-SAD system [10], which has shown consistently good performance in severely noise-corrupted speech. Hence, the combination of SAD (C) to our baseline system (A+B) gives us the SAD-enabled WCE system (A+B+C).

The addition of SAD to our baseline system had a positive impact towards driving the error rates down. However, we continued to observe a large number of false-alarms being generated from speech-only regions. Ideally, the syllable detector generates hits for sonorants in clean speech. However, for noisy speech, the corruption of non-sonorant regions with harmonic noise leads to errors. In order to remove noise harmonics from the noisy speech signal, we decided to investigate several speech enhancement techniques. A key step in all speech enhancement algorithms is a good initial estimate of the noise spectrum. We used the decisions of the Combo-SAD algorithm to obtain this initial noise estimate. By averaging over the local noise frames (as determined by Combo-SAD decisions), we obtained a relatively robust initial estimate of the noise spectrum. Hence, the combination of speech enhancement (D) and SAD (C) to our baseline (A+B) gives us the final proposed WCE system. Figure 3 shows all the major components along with the data flow for the proposed system.

4. Results and Discussion

In this study, we have evaluated the proposed WCE system on Prof-Life-Log as well as Hub-5 evaluation dataset (Hub-5 is a publicly available conversational telephony task and is derived from Switchboard and Call-Home English corpora). In order to measure system performance, we define a very simple metric known as word count error \( WC_{err} \), which is given by:

\[
WC_{err} = 100 \times \frac{|\text{Actual words} - \text{Estimated words}|}{\text{Actual words}}
\]

It is useful to note that the evaluation metric penalizes over and under estimation equally.

The Combo-SAD threshold weight parameter was set to 0.5 for all experiments. We explored a variety of speech enhancement techniques, and found that spectral subtraction (SS) [14] was the most consistent performer. The update parameter for SS was set to 0.05.

In the first experiment, we ran the baseline system on Prof-Life-Log and Hub-5 tasks. For Hub-5 and Prof-Life-Log corpora, the LMMSE estimator was trained on a subset of the Switchboard corpus. Table 1 shows the results for both corpora in terms of word count error. For Prof-Life-Log, the performance (~43%) is much worse than Hub-5 (~11%). As mentioned earlier, this large performance gap is mostly due to a wider variety of noise-types and SNRs in Prof-Life-Log.

Table 1: Comparing performance of Baseline System (A+B) for Hub-5 and Prof-Life-Log datasets.

<table>
<thead>
<tr>
<th>Evaluation Dataset</th>
<th>Baseline System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hub-5</td>
<td>11.39</td>
</tr>
<tr>
<td>Prof-Life-Log</td>
<td>42.97</td>
</tr>
</tbody>
</table>

In the next experiment, we extracted multiple pure-samples (speech-free) of the dominant noise-types from Prof-Life-Log, and used these samples to corrupt the Hub-5 data. This process allows us to study the performance of the proposed method in noise, in a more controlled fashion. Particularly, we selected (i) General-Conference (GC), (ii) Cafeteria, (iii) Office, (iv) Car, and (v) Conference-Banquet noise-types from Prof-Life-Log. These noise-types were chosen since they reflect the diversity of conditions, i.e., from relatively quiet (office) to extremely noisy (conference-banquet), and from babble/large-crowd like (cafeteria) to machine-like (car). Using the FaNT tool [15], we created three SNR variants for each noise-type, i.e., 0dB, 10dB, and 20dB, which gave us a total of 15 variants.
of Hub-5 (5 noise-types and 3 SNR values). We ran the baseline (A+B) as well as the SAD (A+B+C) and Speech Enhancement (A+B+C+D) systems on all 15 variants. The results are shown in Figure 4. The blue, red and black lines represent the performance of baseline, SAD-enabled, and SAD plus enhancement enabled WCE systems, respectively. It can be seen that the addition of SAD and enhancement components progressively drives the word count errors down, for all noise-types and all SNRs. Among the noise-types, the errors are largest and smallest for conference-banquet and office, respectively. On an average, the error rates fall from about 42% to 20% for 0dB SNR, which is remarkable given that the Hub-5 performance for non-noisy data is about 11%.

In the next experiment, we ran the WCE systems on 16 hours of Prof-Life-Log data. The overall as well as environment-specific performance is shown in Figure 5. From the figure, it is seen that conference-banquet and office are the worst and best performing environments (this is similar to what was observed for noisy Hub-5). As seen previously, the error rates progressively decline with the addition of SAD and enhancement components to the WCE solution (the office environment is an exception, where a marginal increase in error is seen with addition of the speech enhancement component). In general, the error rates for Prof-Life-Log are marginally higher than those observed for noisy Hub-5 (at 0dB SNR). The best average performance obtained for Prof-Life-Log is 17% word count error.

In a final experiment, we investigated the trends in cumulative word count estimation error separately for both workdays. The method for obtaining cumulative WCE error is as follows: let \( w_i \) and \( \hat{w}_i \) be the actual and estimated word count for the \( i^{th} \) cut, where \( i = 1, \ldots, N \) and \( N \) is the total number of cuts (and the cuts are arranged chronologically). Now, the cumulative word count error at segment \( P \leq N \) is given by:

\[
\text{Cumulative WCE error} = 100 \times \frac{\sum_{i=1}^{P} w_i - \sum_{i=1}^{P} \hat{w}_i}{\sum_{i=1}^{P} w_i}
\]

By using the time corresponding to the \( i^{th} \) cut, the time evolution of the cumulative word count error can be analyzed.

Figure 6 shows the cumulative word count errors for the two evaluation days from Prof-Life-Log. Here, the results are only shown for the SAD and speech enhancement enabled WCE system. For the typical workday at the University, it can be observed that the estimation error gradually reduces with time. This phenomenon can be attributed to the general canceling of errors where we overestimate the word counts for some cuts, and under-estimate the word counts for others (effectively canceling each other and resulting in lower cumulative error). However, this phenomenon is impacted by noise, as the tendency to over-estimate word count will be higher in noisier conditions (due to relatively larger number of hits during peak-picking). This effect is observed in the trends shown in Figure 6. For example, on the conference workday, the cumulative error continuously diverges when the speaker is present in the general-conference area, walking outdoors towards the conference-banquet, and finally when he is at the conference-banquet (all of these environment are very noisy with cocktail party/large crowd/babble type noise). Therefore, the word count error for the entire UTD workday is much lesser (about 1%) than for Conference workday (about 19%).

5. Conclusion

In this study, we proposed a new syllable based framework for robust word count estimation. We used the syllable rate estimation technique proposed by Wang and Narayanan [1] as our baseline method, and demonstrated its weakness in processing noisy naturalistic long duration audio from Prof-Life-Log. In order to impart robustness to the syllable detector, we propose (i) the use of UTD Combo-SAD technique to eliminate processing of non-speech regions of audio, and (ii) the use of speech enhancement technique to de-noise the spectrum. Both modifications significantly reduce the number of false-alarms in the syllable detector, and improve the accuracy of syllable rate estimation. The proposed solution also employs a simple LMMSE estimator to compute word rate from syllable rate. We evaluated the proposed system on Prof-Life-Log corpus and compared the performance to noisy and clean version of Hub-5 corpus. Our evaluations reveal that the proposed system improves performance across a wide range of noise-types and SNR conditions. The overall solution is computationally inexpensive, and provides word count error rates as low as 1% for relatively quite/noise-free workdays and 19% for very noisy workdays.
6. References


