ROBUST OVERLAPPED SPEECH DETECTION AND ITS APPLICATION IN WORD-COUNT ESTIMATION FOR PROF-LIFE-LOG DATA

Navid Shokouhi, Ali Ziaei, Abhijeet Sangwan, John H. L. Hansen[⋆]

Center for Robust Speech Systems (CRSS)
The University of Texas at Dallas, Richardson, TX 75080-3021, USA
{navid.shokouhi,ali.ziaei,abhijeet.sangwan,john.hansen}@utdallas.edu

ABSTRACT

The ability to estimate the number of words spoken by an individual over a certain period of time is valuable in second language acquisition, healthcare, and assessing language development. However, establishing a robust automatic framework to achieve high accuracy is non-trivial in realistic/naturalistic scenarios due to various factors such as different styles of conversation or types of noise that appear in audio recordings, especially in multi-party conversations. In this study, we propose a noise robust overlapped speech detection algorithm to estimate the likelihood of overlapping speech in a given audio file in the presence of environment noise. This information is embedded into a word-count estimator, which uses a linear minimum mean square estimator (LMMSE) to predict the number of words from the syllable rate. Syllables are detected using a modified version of the *mrate* algorithm. The proposed word-count estimator is tested on long duration files from the Prof-Life-Log corpus. Data is recorded using a LENA recording device, worn by a primary speaker in various environments and under different noise conditions. The overlap detection system significantly outperforms baseline performance in noisy conditions. Furthermore, applying overlap detection results to word-count estimation achieves 35% relative improvement over our previous efforts, which included speech enhancement using spectral subtraction and silence removal.

Index Terms— Word-count estimation, overlapped speech detection, Massive audio data, Prof-Life-Log

1. INTRODUCTION

The ability to automatically count the number of words spoken by an individual over a certain period of time (word-count estimation) is important in a number of fields. Large scale word-count estimation (i.e. over long durations) is beneficial in determining the amount to which a child is exposed to new words; a factor that has proven essential for language acquisition and development [1]. In [1], Hart et al. showed the high correlation between more advanced language abilities and academic success observed in children that are exposed to greater word-count rates in early stages of language development. The relationship between word-count values and early signs of autism has also been the subject of studies [2]. Word-count values are also useful in the analysis of massive audio data, such as the Prof-life-log corpus [3]. Prof-life-log is a speech corpus that contains long durations of audio recordings. In this collection, the primary speaker wears a portable LENA recording device [4] throughout the workday. Although the primary speaker is always the same individual, the people he interacts with vary frequently. The LENA unit is small in dimension and causes minimal self-awareness for the primary speaker and people with whom he interacts, allowing recordings to capture realistic conversations. Recording durations typically vary between 6-to-8 hours and take place in various environments and noise conditions. An estimate of the primary speaker‘s speech activity (i.e. word-count) facilitates analyses that predict the level of productivity and helps determine areas in the recording that are more valuable for further processing. Prof-Life-Log is the context in which we intend to investigate the performance of our proposed word-count estimator (WCE).

One can estimate word-count by either (1) performing automatic speech recognition (ASR), or (2) take an indirect approach using acoustic characteristics of the signal to detect syllables [5] and thereby estimate the word rate. Previously in [6], Ziaei et al. developed a WCE for Prof-Life-Log in which they used the latter approach. The proposed WCE was a linear minimum mean square estimator (LMMSE) that mapped syllable rates to word-count rates. There, the effect of noise on word-count accuracy was observed and accounted for by applying 1) spectral subtraction, to enhance the speech, and 2) silence removal, to reduce the false alarm rates in the syllable detector. Despite those efforts, the problem of secondary speakers, especially in crowded environments, was still significant. In segments detected as speech for the primary speaker, if secondary speakers are also talking (typical speech activity detection algorithms are blind to the number

⋆This project was funded by AFRL under contract FA8750-12-1-0188 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

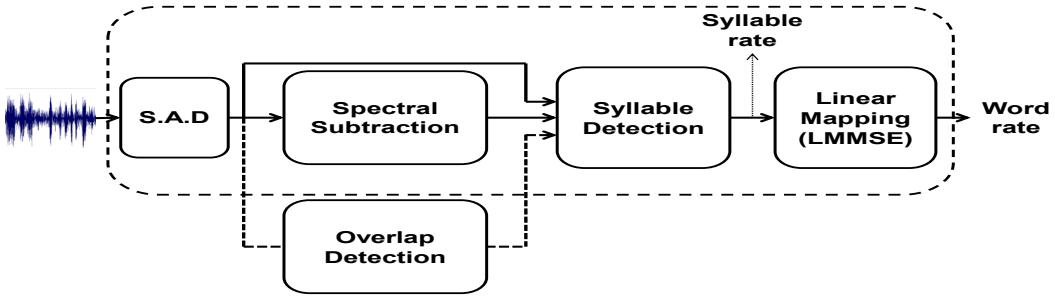

Fig. 1. Word-count estimation system configuration. The overlap detection system is shown as an addition to the original system.

of speakers), the WCE system overestimates the rate of voiced syllables. When there is a clear boundary between the primary and secondary speaker's speech, one can remove the interfering speech by merely applying an energy threshold; since the primary speaker is much closer to the device. In the case of overlapping speech, however, extra syllables are inserted in between the primary speaker's speech and removing such instances requires more sophisticated processing.

In this study, our goal is to formulate a method in which we can account for overlapped speech in noisy audio recordings by detecting such regions. Despite numerous studies targetting overlapped speech detection [7, 8, 9, 10], overlap detection in the presence of noise has seldom been visited [11]. Through this, we improve the accuracy of our word-count estimation algorithm by removing overlapped regions, which are responsible for most of the insertion errors syllable detection [6]. We start by introducing the WCE system configuration, describing where we intend to embed the overlap detection system, section 2. Section 3 illustrates the proposed overlap detection algorithm. System performances are demonstrated in section 4 followed by conclusions.

2. WORD-COUNT ESTIMATION

The word-count estimator proposed in [6] estimates the number of words per unit time by applying a linear transformation to the syllable rate. Syllable rates are calculated based on a modified version of the *mrate* algorithm [5] using acoustic characteristics of the signal: pitch, smoothed spectrogram. For a detailed description of the syllable detection algorithm see [5, 6]. This algorithm detects the location of syllables in a given speech segment, which is then used to calculate syllable rates. It is shown in [6] that with the help of a linear minimum mean square estimator (LMMSE), linear coefficients can be derived that map the syllable rates to the number of words per unit time.

$$\tilde{\mathbf{a}} = argmin_{\mathbf{a}}\Big\{\frac{1}{N}\sum_{\mathbf{a}}\big(W_r(n) - \mathbf{a}S_r(n)\big)^2\Big\}, \qquad (1)$$

where $W_r(.)$ and $S_r(.)$ are the word-count and syllable rates at any given time, respectively. n indicates the index of a given time segment and N is the total number of segments used to train the linear transformation parameter, $\tilde{\mathbf{a}}$. The linear transformation parameter(s)[1] can be trained using background conversational data that include transcriptions, so that the target word-counts are also avaible. For this study, we rely on a subset of Prof-Life-Log data that has been manually transcribed.

In [6], higher accuracy is obtained by introducing speech activity detection (SAD) [12] and spectral subtraction to the front-end of the WCE. SAD reduces false alarms by omitting non-speech regions, which helps avoid detection errors by the syllable detector. Spectral subtraction enhances speech regions, allowing the syllable detector to detect voiced regions more accurately. None of these techniques, however, are able to address the issue of overlapped speech.

2.1. Incorporating Overlap Detection in Word-Count Estimation

An estimation of the location and amount of overlapped speech in a given speech segment can be combined with SAD labels to supply an additional layer of data pruning before syllable detection. Figure 1, shows the proposed WCE system configuration. Initially, SAD is performed on the raw data to detect speech locations. From SAD labels, the non-speech regions are used to estimate the noise level in each short segment and submitted to the spectral subtraction algorithm. The speech-only segments are passed to the syllable detector after applying spectral subtraction. Finally, syllable rates (calculated by dividing the number of syllables by the segment length) are transformed into word-count rates using LMMSE coefficients. In our proposed system, overlap detection outputs are combined with SAD results, to provide an extra layer of data pruning (as seen in Fig. 1).

[1]Note that $\tilde{\mathbf{a}}$ is in general a vector parameter comprised of a bias factor and a linear coefficient. In cases where the bias factor is used, $S_r(n)$ is replaced by $\begin{bmatrix} S_r(n) & 1 \end{bmatrix}^T$

3. OVERLAPPED-SPEECH DETECTION

Detecting regions of overlapped speech has proven to be useful in applications such as speaker identification (SID) and speaker diarization [9]. In all these applications, the presence of a secondary speaker either decreases model reliabilities (in training), or introduces confusion in the scoring process by distorting test files.

When interfering speakers speak at the same time as the primary speaker (i.e., overlapping speech), removing their speech requires sophisticated processing; hence, we rely on detecting these regions. One of the setbacks in detecting overlapped speech in Prof-Life-Log is the high amount of noise, which makes using traditional overlap detection methods [13, 14] less functional (see section 4.1). Hence, we propose a novel approach for overlap detection based on enhanced spectrograms. These enhanced spectrograms, called *pyknograms*, were first introduced in [15] to facilitate formant tracking and are calculated by applying multiband demodulation in the framework of the AM-FM modulaton model [16]. We take advantage of this approach to enhance noisy spectrograms and develop our overlap detection algorithm. The next section briefly describes the algorithm through which pyknograms are obtained. Readers are encouraged to visit [15] for a more detailed discription.

3.1. Extracting pyknograms

In pyknograms, the resonances (formants) and harmonic structure of speech are enhanced by decomposing the spectral sub-bands into amplitude and frequency components. This multiband analysis uses the AM-FM speech model [16] to decompose the subbands and thereby calculate their corresponding instantaneous frequencies and bandwidths. The speech signal is passed through a gammatone filterbank (our studies show that using logarithmically spaced gammatone filters is more effective in capturing harmonic structures, while [15] uses linearly spaced Gabor filters [2]). Each resulting subband is then decomposed into amplitude and frequency components using the discrete energy separation algorithm (DESA-1) [16], where the frequency and amplitude components of a given subband, $x(n)$, are calculated using the discrete energy operator,

$$\Psi[x(n)] = x^2(n) - x(n-1)x(n+1), \tag{2}$$

with the following equations.

$$f = \frac{1}{2\pi} \arccos\left(1 - \frac{\Psi[x(n) - x(n-1)]}{2\Psi[x(n)]}\right) \tag{3}$$

$$|a| = \sqrt{\frac{\Psi[x(n)]}{\sin^2(2\pi f)}} \tag{4}$$

[2]It is worth mentioning that the goal in [15] was to detect formant locations, whereas in this study we intend to enhance prominent harmonics.

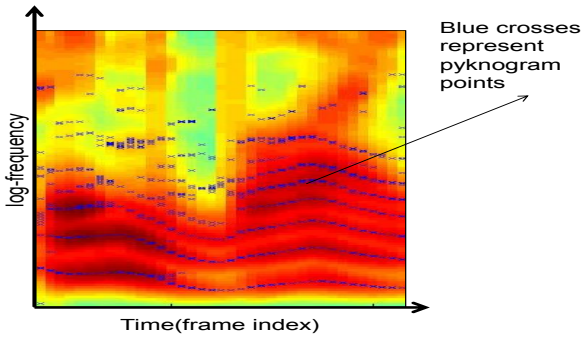

Fig. 2. Demonstration of a pyknogram. The blue crosses show locations of the pyknogram points on the time-frequency scale. The spectrogram of the corresponding speech segment is plotted in the background for comparison.

The weighted average of the instantaneous frequency components are used to derive a short-time estimate value for the dominant frequency in each subband over a fixed period of time, in this case the duration of a time-frame (typically 25 msec).

$$F_w(t) = \frac{\sum_t^{n+T} f(n)a^2(n)}{\sum_t^{n+T} a^2(n)}, \tag{5}$$

where $f(n)$ and $a(n)$ are the instantaneous frequency and amplitude functions calculated for each sample in the t^{th} frame over the frame length (T samples per frame). Resonances and harmonic peaks are located in each frame by comparing the average frequency estimates with filterbank center frequencies [15].

The motivation behind using an energy operator based approach [16] is to avoid assumptions on the number of speakers in the signal. The AM-FM decomposition method relies on signal resonances and does not restrict the signal to a specific structure. The final time-frequency representation is called a pyknogram and is denoted $S_{pyk}(t, f)$ as a function of time (t) and frequency (f).

3.2. Detecting overlaps from pyknograms

As a final step in detecting overlapped regions, we use sudden jumps in the harmonic structure as an indication of interfering speech. One can use the analogy that speech harmonic patterns resemble skiing tracks. In the case of a single speaker, the patterns form parallel tracks progressing over time. In the presence of an interfering speaker, these patterns are distorted by similar, but intersecting tracks, which increases the time difference between the patterns. We use the difference between adjacent frames as our measure of overlapped speech. The distance function, D_{ovl}, at frame t is computed as the Euclidean distance between consecutive pyknogram frames, $S_{pyk}(t, f)$ and $S_{pyk}(t-1, f)$.

$$D_{ovl}(t) = \sqrt{\sum_{f} \Big(\big(S_{pyk}(t,f) - S_{pyk}(t-1,f) \big)^2 \Big)}. \quad (6)$$

Overlapped segments are expected to have greater values as compared to single-speaker speech. We use manually labeled data to find the optimum threshold for D_{ovl}. This threshold is selected to minimize the equal-error-rate of overlap detection.

4. EXPERIMENTS AND RESULTS

In this section we first investigate the performance of our overlap detection algorithm in noise and compare it with a baseline overlap detection system that uses features called Gammatone Sub-band Frequency Modulations (GSFM) [17]. In 4.2, we measure errors in word-count estimatation with and without the use of overlap detection.

4.1. Overlap detection accuracy

Typical overlap detection systems are designed to detect overlapping speech segments in files that are generally clean of any environment noise, which makes them less reliable for data collected in real meeting and conversation scenarios, such as Prof-Life-Log.

In order to evaluate overlap detection performance, our overlap detection experiments are initially conducted on the speech separation challenge (SSC) database [18]. This database provides a manageable set of artificially generated overlapped speech files. Each file is created by summing two utterances spoken by 2 separate speakers from a pool of 34 speakers. We also have access to files that are "clean" of overlapping speech. In order to evaluate the performance in of our overlap detection system we use clean files as target and overlapped files as non-target files (or vice versa). For consistent performance in overlap detection, we use overlapped files with average signal-to-interference (SIR) of $0dB$, which means that the two utterances are mixed with the same average energy. The SIR value is a key component in overlapped speech detection performance [17]. We use equal error rates (EER, when false alarms and missed rates are equal) as our measure of system performance. To measure performance under noise, files are mixed with noise samples extracted from Prof-Life-Log recordings with SNR values ranging from clean($100dB$) to $-10dB$. It is important in this context that the difference between SIR and SNR be clear to the reader. SNR specifies the amount of noise added to the files (overlapped or not) and SIR determines the relative energy of the two utterances in overlapped files. The noise used here is the same in our word-count experiment. Figure 3 shows overlap detection EER values for different SNR values and compares the performance with GSFM features. As seen in the figure, GSFM performance drops dramatically even for the most trivial noisy condition ($20dB$).

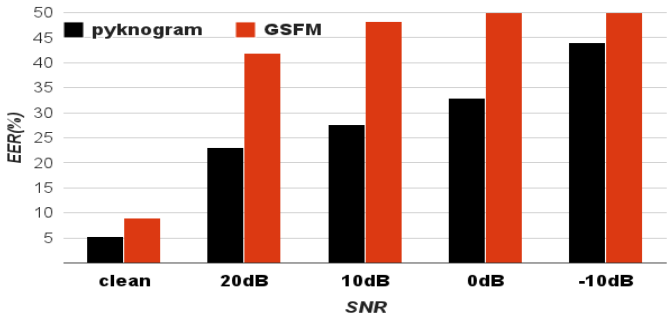

Fig. 3. A comparison of overlap detection equal error rates (EER) for pyknogram (proposed) and GSFM-based systems for different amounts of added noise. It is clear that GSFM is vulnerable even to the slightest amount of noise (high SNR).

Table 1. WCE performance in Prof-Life-Log with respect to overlapped speech.

$\frac{minimum\ mean\ square\ Error}{\#of words}$	
overlaps NOT removed	5.71%
overlaps removed	**3.69%**

It is worth mentioning that we did not include performances from other overlap detection algorithms, since to the best of our knowledge none of the existing algorithms claim robustness in noisy conditions.

4.2. Word-count estimation experiments

Word rates are extracted from 5 days of prof-life-log recordings. Each day contains roughly 6 to 8 hours of audio which has been transcribed to include the transcriptions of the primary speaker's speech, speaker labels (primary vs. secondary), and the type of environment in which the recordings take place. We have mostly concentrated on environments that are more likely to contain overlapping speech, such as multi-party meetings and conferences. The recording sampling frequency from the LENA device is $44.1kHz$, which we have down-sampled to $8kHz$. Table 1 shows over 35% improvement in relative mean square error after removing overlapped regions.

5. CONCLUSIONS

In this study we proposed a novel approach for noisy overlapped speech detection. In addition, by combining the results of the overlap detection algorithm with an existing word rate estimation algorithm, we were able to decrease relative mean square errors by 35%. The proposed word-rate estimator presents a valuable tool in processing massive audio data.

6. REFERENCES

[1] B. Hart and T. R. Risley, *Meaningful differences in the everyday experience of young American children*, Brookes Publishing, 1995.

[2] D. Xu, J. Gilkerson, J. Richards, U. Yapanel, and S. Gray, "Child vocalization composition as discriminant information for automatic autism detection," in *Engineering in Medicine and Biology Society, EMBC*, September 2009, pp. 2518–2522.

[3] A. Ziaei, A. Sangwan, and J. H. L. Hansen, "Prof-life-log: Personal interaction analysis for naturalistic audio streams," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 7770–7774.

[4] X. Dongxin, U.H. Yapanel, S.S. Gray, J. Gilkerson, J.A. Richards, and J.H.L. Hansen, "Signal processing for young child speech language development," in *Workshop on Child Computer Interaction (WOCCI)*, 2008.

[5] D. Wang and S. Narayanan, "Robust speech rate estimation for spontaneous speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15.8, pp. 2190–2201, 2007.

[6] A. Ziaei, A. Sangwan, and J. H. L. Hansen, "A speech system for estimating daily word counts," in *Proc. INTERSPEECH*, Singapore, September 2014.

[7] B.Y. Smolenski and R.P. Ramachandran, "Usable speech processing: A filterless approach in the presence of interference," *Circuits and Systems Magazine, IEEE*, vol. 11, no. 2, pp. 8 –22, 2011.

[8] K. Krishnamachari, R. E. Yantorno, D. S. Benincasa, and S. J. Wenndt, "Spectral autocorrelation ratio as a usability measure of speech segments under co-channel conditions," in *IEEE Intl. Symp. on Intelligent Signal Processing and Communication Systems, ISPACS*, November 2000, pp. 710–713.

[9] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, "Overlapped speech detection for improved diarization in multiparty meetings," in *Proc. ICASSP*, Las Vegas, Nevada, 2008, pp. 4353–4356.

[10] K. Krishnamachari, R. E. Yantorno, J. M. Lovekin, D. S. Benincasa, and S. J. Wenndt, "Use of local kurtosis measure for spotting usable speech segments in co-channel speech," in *Proc. IEEE ICASSP*, Salt Lake City, Utah, 2001, pp. 649–652.

[11] N. Shokouhi, A. Sathyanarayana, S.O. Sadjadi, and J. H. L. Hansen, "Overlapped-speech detection with applications to driver assessment for in-vehicle active safety systems," in *Proc. IEEE ICASSP*, Vancouver, BC, May 2013.

[12] S. O. Sadjadi and J. H. L. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Process. Lett.*, vol. 20, pp. 197–200, March.

[13] N. Shokouhi, A. Sathyanarayana, S. O. Sadjadi, and J. H. L. Hansen, "Overlapped-speech detection with applications to driver assessment for in-vehicle active safety systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 2834–2838.

[14] K. Boakye, *Audio Segmentation for Meeting Speech Processing*, Ph.D. thesis, Fall 2008.

[15] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *The Journal of the Acoustical Society of America*, vol. 99.6, pp. 3795–3806, 1996.

[16] P. Maragos, Kaiser J. F., and Quatieri T. F., "Energy separation in signal modulations with application to speech analysis," *IEEE Transactions on Signal Processing*, vol. 41.10, pp. 3024–3051, 1993.

[17] N. Shokouhi, S. O. Sadjadi, and J. H. L. Hansen, "Co-channel speech detection via spectral analysis of frequency modulated sub-bands," in *Proc. INTERSPEECH*, Singapore, September 2014.

[18] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, November 2006.