ProfLifeLog: ENVIRONMENTAL ANALYSIS AND KEYWORD RECOGNITION FOR NATURALISTIC DAILY AUDIO STREAMS

Abhijeet Sangwan, Ali Ziaei, and John H. L. Hansen

Center for Robust Speech Systems (CRSS),
Department of Electrical Engineering, University of Texas at Dallas, Richardson, Texas, U.S.A

ABSTRACT

This study presents keyword recognition evaluation on a new corpus named ProfLifeLog. ProfLifeLog is a collection of data captured on a portable audio recording device called the LENA unit. Each session in ProfLifeLog consists of 10+ hours of continuous audio recording that captures the work day of the speaker (person wearing the LENA unit). This study presents keyword spotting evaluation on the ProfLifeLog corpus using the PCN-KWS (phone confusion network-keyword spotting) algorithm [2]. The ProfLifeLog corpus contains speech data in a variety of noise backgrounds which is challenging for keyword recognition. In order to improve keyword recognition, this study also develops a front-end environment estimation strategy that uses the knowledge of speech-pause decisions and SNR (signal-to-noise ratio) to provide noise robustness. The combination of the PCN-KWS and the proposed front-end technique is evaluated on 1 hour of ProfLifeLog corpus. Our evaluation experiments demonstrate the effectiveness of the proposed technique as the number of false alarms in keyword recognition are reduced considerably.

Index Terms— Keyword Spotting, Phone Confusion Networks, Environment Estimation, False Alarms, Noise Robustness

1. INTRODUCTION

In this study, we introduce the ProfLifeLog corpus which is being collected with an intention of developing and evaluating speech systems on data captured in natural settings. The ProfLifeLog corpus uses the LENA unit which is a portable device that can capture up to 10+ hours of audio recording in a single session. The most popular use of the device has been to capture the language environments of infants and young children, where the subject in question wears the unit. Subsequently, speech processing software has been used to analyze the collected data for various metrics of interest such as adult word count, adult-child turn-taking count, child vocalization count, TV (television) time *etc.* [1]. More recently, there has been an interest in using the device with older children and adults with focus on studying language acquisition in bilingual environments, vocabulary and accent tracking in second language speakers, *etc.*

Data analysis for LENA would be strengthened by the use of Keyword Spotting (KWS) technology. Since LENA recordings are collected in real-world conditions, noisy audio also represents a major challenge for keyword recognition. This study examines the effectiveness of the PCN-KWS (phone confusion network-keyword spotting) algorithm for keyword recognition on ProfLifeLog data. PCN-KWS is a new algorithm for Keyword Spotting (KWS) that is based on searching keywords in phone confusion networks (PCNs) [2]. The PCN-KWS algorithm has shown good recall and precision performance on keyword spotting tasks when applied to standard speech corpora such as Switchboard, SPINE (Speech In Noise corpus), and TIMIT. Motivated by these results, this study examines the performance of the KWS-PCN algorithm on ProfLifeLog corpus data which presents more challenging speech data.

Phone-based KWS algorithms are inherently prone to higher rates of false-alarms, especially in noisy conditions. This is especially true in the ProfLifeLog corpus where speech data is captured in a variety of real world interactions such as buying a sandwich, calling while driving, *etc*. In order to mitigate the impact of false alarms on keyword recognition performance, we propose an efficient front-end environment estimation scheme which utilizes the knowledge of signal to noise ratio (SNR), and speech-pause decisions (voice activity detection) to reduce the false-alarm rate in keyword detection. The combination of the proposed front-end environment estimation with the PCN-KWS algorithm lends noise robustness to the keyword recognition system.

2. PROFLIFELOG CORPUS

The ProfLifeLog corpus is a database of speech material collected on the LENA unit. LENA stands for Language Environment Analysis [10]. It has been primarily used for analyzing the language environment of infants. The LENA unit has a small form factor and is light enough to be carried in a shirt pocket (see Fig. 1). In our data collection, the unit is attached to a person who then carries the unit for the entire work day. As a result, the device captures all speech and background data as the speaker performs his day-to-day tasks. As a result, the collected data contains speech in a variety of background noise types ranging from very quiet (such as office) to very noisy (such as restaurant). In general, the audio material has been collected in natural settings where the user activity is not controlled. Hence, the data collected offers an excellent opportunity to evaluate speech systems on real world data. So far, the ProfLifeLog corpus contains 30 days of audio recording resulting in a total collection of 300+ hours. We have also initiated the transcription effort for ProfLifeLog, and used 1 hour of transcribed data for evaluation in this study. For evaluating the proposed front-end scheme in conjunction with the PCN-KWS algorithm, we extracted two 30 minute speech segments from the ProfLifeLog corpus. While one segment was recorded in a meeting-environment, the other was collected in restaurant-environment.

This project was funded by The University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. H. L. Hansen.

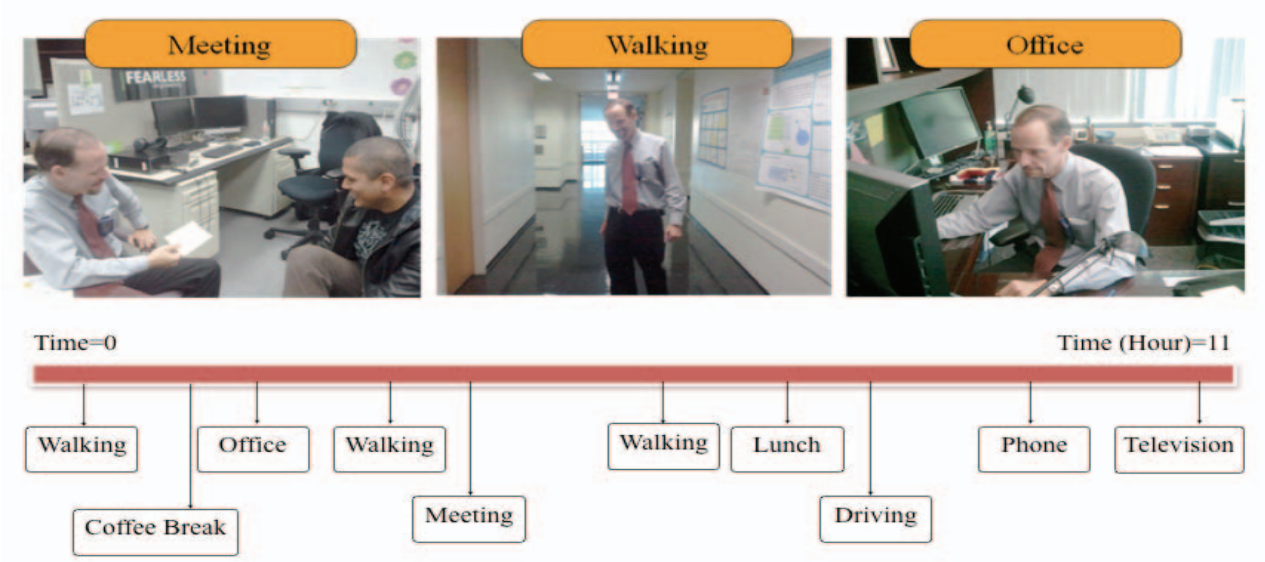

Fig. 1: Data collection using the LENA unit: A single session consists of 10+ hours of audio recording with the speaker constantly carrying the unit. Speech is collected in a wide variety of backgrounds such as Restaurant, Office, Meeting, Walking, Driving *etc.*

3. PHONE CONFUSION NETWORK BASED KWS

In this section, we briefly review the PCN-KWS algorithm. In the PCN-KWS algorithm, the speech signal is first decoded using monophone HMMs, and the corresponding phone lattices are generated. In the next step, the phone lattices are converted into phone confusion networks (PCNs). Finally, the PCN-KWS algorithm is used to search the PCNs for keywords.

It is noted that KWS algorithms based on phone lattices have been proposed in the past [3, 4, 5]. The general approach is to search for the phone sequence corresponding to the keyword in the lattice. In such approaches, the likelihood of the phone sequence in the lattice can be compared to a threshold to make a decision. Additionally, phone substitution, insertion, and deletion rules can also be used to account for errors in phone decoding. For example, phone confusion matrices generated by comparing ASR output with ground truth transcriptions have been used to objectively compute the likelihood of phone substitution, deletion, and insertion errors [6]. Subsequently, the phone confusion matrices are used to re-estimate the likelihood of phone sequences in the lattice. However, ASR lattices are generally large and searching them can be time consuming. On the other hand, confusion networks (CNs) are a more compact form of speech recognition lattices, and have also been shown to deliver lower WERs (word error rates) in ASR (automatic speech recognition) tasks [7]. These properties of CNs make them suitable for fast and accurate keyword searching.

The PCN-KWS algorithm searches for the phone sequence corresponding to the keyword in the phone confusion network. In particular, it uses the Viterbi algorithm to find the most likely occurrence of the keyword within a PCN. However, since ASR phone-decoding is inherently imperfect, PCNs contain a number of substitution, insertion, and deletion errors. Therefore, a simple strategy such as searching for the phone-sequence corresponding to the keyword in the PCN realizations will be error prone. While substitution errors are easily handled within the PCN structure (by choosing the desirable phone-sequence realization among alternatives), the PCN-KWS provides additional consideration for handling insertion and deletion errors. In particular, insertion errors are mitigated by re-interpreting the probability of *e* (special empty node in CNs) as the probability of self transition. The PCN-KWS algorithm also considers the timing information of nodes while searching for valid paths within the PCN. Finally, the algorithm also allows for phone deletion by introducing a deletion penalty. More details of the algorithm can be found in [2].

4. FRONT-END ENVIRONMENT ESTIMATION

The proposed front-end environment estimation combines two voice activity detection (VAD) techniques, namely, MO-LRT VAD system proposed by Ramirez [8] and a standard HMM (Hidden Markov Model) based phone decoder. In our experimental studies, we have observed that a fusion strategy that combines MO-LRT and phone-decoder based VADs is able to deliver higher performance (up to 10% improvement in accuracy). This is the motivation behind using a fusion strategy for VAD. The process of combining VAD decisions is illustrated in Fig. 2. The two VAD systems work independently on the data to produce the speech and pause likelihoods for each frame. As shown in the figure, the MO-LRT generates the speech and pause conditional likelihoods for the i^{th} frame, *i.e.*, $L^R(s_i)$ and $L^R(n_i)$, respectively. In parallel, the phone decoder is used to generate time-aligned phone transcripts for the speech signal. Subsequently, as shown in Fig. 2 the speech likelihoods for the vowel frames ($L^H(s_i)$) are assigned the normalized acoustic scores. On the other hand, the noise likelihood for the vowel frames is assigned 0. This process is motivated by the capability of the phone decoder to detect vowel regions in speech with high confidence. Similarly, the noise likelihoods for noise-only frames $L^H(n_i)$ are assigned the normalized acoustic scores, and the speech likelihoods are assigned 0. For all other phone types (*i.e.*, stops, nasals, fricatives, semivowels), the noise and speech likelihoods are assigned as 0. This reflects the ambiguity in phone decoder output where non-vowel regions may not be detected with high accuracy. For the final decision, a combined speech and noise likelihood is generated by adding the MO-LRT speech and noise likelihoods, with HMM-based speech and pause likelihoods, respectively. In this manner, the HMM gen-

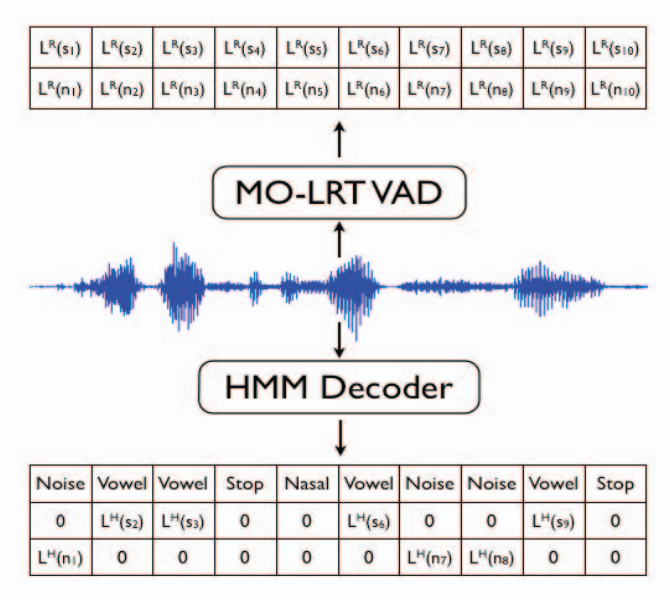

Fig. 2: Proposed Fusion based VAD System.

erated likelihoods serve to bias the MO-LRT likelihood whenever speech or noise is detected with high confidence. Thereafter, the log likelihood ratio (LLR) is computed by using the combined speech and noise likelihoods, and compared to a threshold for speech-pause decisions. In this study, the HMM based phone decoder was trained on broadcast news (BN) corpus.

It is important to mention that the normalized log-likelihood obtained by HMM decoder for a vowel part will be assigned to each frame in that part equally. At last normalized log-likelihood obtained by HMM decoder and MO-LRT VAD will be added together and the final decision will be taken based on assigned threshold. The same procedure is used for frames detected as noise.

Once the speech-pause decision are made, the SNR (signal-to-noise ratio) estimate for each frame is computed as follows. Let $Y(k,i)$ be the k^{th} Fourier transform coefficient for the i^{th} frame, and let $\lambda_n(k,i)$ be the estimate of the noise power spectrum. Here, $\lambda_n(k,i)$ has been obtained by following the process described in [9]. Now, the SNR estimate for the k^{th} coefficient and i^{th} frame can be obtained as

$$\gamma(k,i) = \frac{|Y(k,i)|^2}{\lambda_n(k,i)}. \quad (1)$$

The SNR estimate for the frame is obtained by averaging over all frequency bins.

In this manner, the proposed environment estimation front-end generates speech-pause decisions as well as SNR estimates at a frame level. The speech-pause decisions from the environment estimation scheme are used to segment the signal into noisy-speech and noise-only regions. Subsequently, the noise only regions are discarded and not processed by the PCN-KWS algorithm. Furthermore, the SNR estimates for speech segments are now obtained by averaging the SNR values of all frames. Speech segments with low SNR values tend to have poorer recognition accuracy, and hence are eliminated from further processing. In summary, the proposed front-end technique acts as a filter, and retains only relatively clean speech (high SNR) for keyword spotting. Therefore, the proposed strategy should be beneficial towards lowering false-alarm rates.

5. EXPERIMENTS AND RESULTS

The ASR system used as part of the keyword recognition experiment was trained on BN (broadcast news) corpus. The BN data was used to train 128-mixture monophone HMMs. The same acoustic model was also used for the HMM-decoder based VAD. Additionally, a bigram phonotactic language model (LM) was also trained using the BN corpus text transcriptions. Here, the canonical pronunciations were used to convert words into phone sequences. The SPHINX recognition engine was employed to generate phone lattices, and the SRILM toolbox was used to generate the PCNs [7].

In order to evaluate the performance of our proposed VAD system, we synthesize 30 minutes of noisy data by combining clean speech (collected on LENA) with noise recordings of meeting and restaurant environments (also collected on LENA). The noise and speech signals are added to produce noisy speech at 0dB, 10dB, and 20dB SNR. We preferred this process as the ground truth annotations for clean speech was easily available, and noise could be added while controlling the SNR. As a result, the evaluation is more objective and comparable to others reported in literature. It is noted that for KWS evaluation we have not synthesized data but used the recordings directly from LENA.

Figure 3 shows the ROC (receiver operating characteristic) curves for the proposed VAD system in meetings and restaurant environments. It is observed that at 20dB SNR, the VAD performances in both restaurant and meetings environment are comparable and very high (approximately 95% accuracy). Additionally, a drop in performance is observed with decreasing SNR for both environments. Finally, the VAD performance drops more significantly in the restaurant environment. This is expected as the restaurant environment is more challenging.

In order to allow comparison between the PCN-KWS algorithm and other KWS algorithms, we first present the evaluation of the PCN-KWS system on the TIMIT corpus. For this experiments, we followed a keyword recognition setup very similar to that presented in [3]. We chose 200 unique keywords of 6-phone length from the TIMIT test corpus, with a total of 644 occurrences in the database. We eliminated sa1 and sa2 sentences from our evaluation. Figure 4 plots the average false-alarms per keyword against miss-rate for the TIMIT evaluation. The evaluations results show very low false-alarm rates (<11 average false alarm per keyword) for low miss-detection rates (<9%). In fact, the results obtained in this experiment are comparable to the evaluation results in [3]. Additionally, the PCN-KWS system has been evaluated on Switchboard and SPINE corpora, and the details of the performance can be found in [2].

For evaluation of ProfLifeLog data, we chose 20 unique keywords with a total of 153 occurrences. The total amount of audio material used was 1 hour long, and was chosen in equal proportion from two different environments, namely, meetings and restaurant. The restaurant environment was relatively more noisy, and consisted of babble and cocktail party noises. The meetings environment was relatively cleaner. Figure 4 shows the keyword recognition performances in restaurant and meetings environment, respectively. From Fig. 4, it is observed that the keyword recognition performance on ProfLifeLog data is lower than TIMIT data. The reduced performance reflects the challenges in ProfLifeLog data, *i.e.*, spontaneous speech in natural settings. Additionally, it is observed that the keyword recognition performance in meetings environment is superior to that in restaurant environment. Additionally, the use of environment estimation in both the meetings and restaurant data is beneficial towards improving keyword recognition accuracy. From data analysis, it was discovered that in the meetings data, the environment

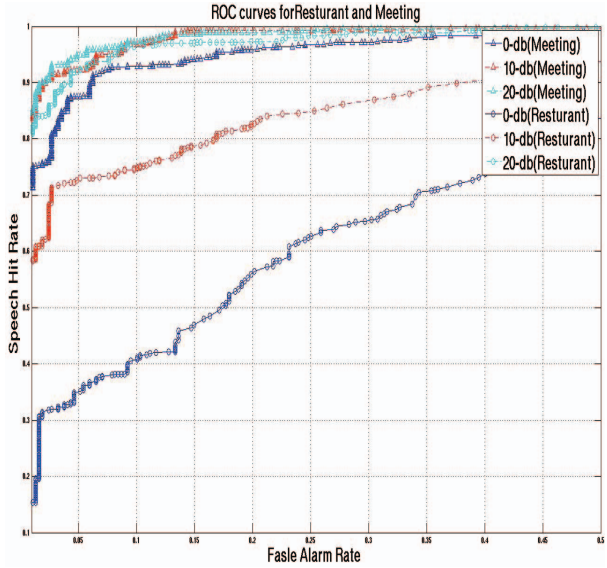

Fig. 3: Proposed VAD performance on ProfLifeLog in meetings and restaurant environments and 0dB, 10dB, and 20dB SNRs.

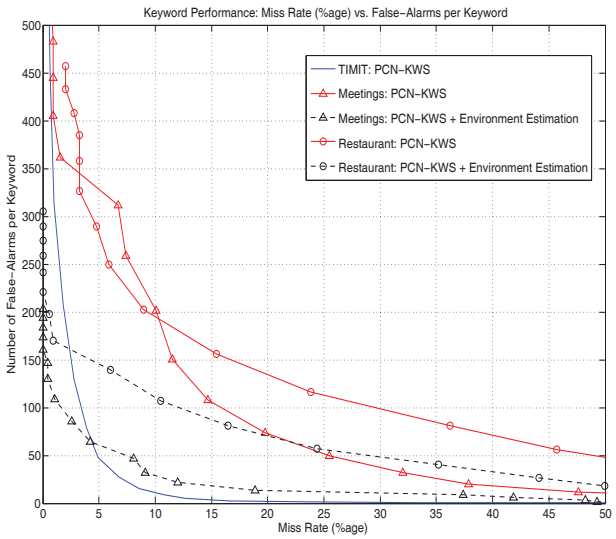

Fig. 4: Keyword Recognition Performance for LENA data: Average Number of False-Alarms per Keyword *vs.* Miss Rate in restaurant and meetings environment. Performance on TIMIT corpora is also shown for comparison.

estimation algorithm is able to separate pauses as well as speech of secondary speaker (person talking to the speaker wearing the LENA unit). Here, eliminating secondary speaker data as well as background leads to improved keyword recognition accuracy. A similar effect was also observed in restaurant data. However, the effectiveness of separating background and secondary speaker from primary speaker is diminished in restaurant setting owing to higher background noise. Therefore, environment estimation benefits meetings more than restaurant. Overall, by using the environment estimation scheme we can move KWS accuracy in meetings environments to TIMIT-like performance, but more work is required to improve performance in restaurant environment (or non-stationary low SNR noise environments in general).

6. CONCLUSION

A new corpus called the ProfLifeLog corpus has been presented. In this new corpus, speech data has been captured using a portable audio recording device called the LENA unit. ProfLifeLog corpus contains long continuous recordings (10+ hours) per session where the day-to-day activity of speakers is captured. This study has also presented keyword recognition evaluation on the ProfLifeLog corpus. Particularly, the study has proposed a new front-end environment estimation algorithm that can generate speech-pause decisions along with SNR (signal to noise ratio) estimates. It has been shown that when this new technique is used in conjunction with the PCN-KWS (phone confusion network-keyword spotting) algorithm, the number of keyword false-alarms can be dramatically reduced. The study presents the feasibility of extracting useful information from long duration collections which may open new avenues in speech and language research.

7. REFERENCES

[1] D. Xu, U. Yapanel, S. Gray, J. Gilkerson, J. Richards, and J.H.L. Hansen, "Signal processing for young child speech language development," in *1st Workshop on Child, Computer and Interaction*, 2008.

[2] A. Sangwan and J.H.L. Hansen, "Keyword recognition with phone confusion networks and phonological features based keyword threshold detection," in *40th Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, pp. 711–715, Nov. 2010.

[3] K. Thambiratnam and S. Sridharan, "Dynamic match phone-lattice searches for very fast and accurate unrestricted vocabulary keyword spotting," in *ICASSP*, March 2005, pp. 465 – 468.

[4] K. Iwata, K. Shinoda, and S. Furui, "Robust spoken term detection using combination of phone-based and word-based recognition," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[5] K. Audhkhasi and A. Verma, "Keyword search using modified minimum edit distance measure," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. IEEE, 2007, vol. 4, pp. 926–929.

[6] P. Zhang, J. Shao, J. Han, Z. Liu, and Y. Yan, "Keyword spotting based on phoneme confusion matrix," in *Proc. ISCSLP 2006*, vol. 2, pp. 408–419.

[7] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," *Computer Speech and Language*, vol. 14, no. 4, pp. 373–400, 2000.

[8] J. M. Górriz, J. Ramírez, J. C. Segura, and C. G. Puntonet, "An improved mo-lrt vad based on a bispectra gaussian model," *IEEE Signal Processing Letters*, vol. 41, no. 15, pp. 877–879, 2005.

[9] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. On Speech and Audio Signal Processing*, vol. 11, no. 5, pp. 466–475, s 2003.

[10] http://www.lenafoundation.org