ACOUSTICS 2012

Noise-robust speech recognition system based on power spectral subtraction with a geometric approach

J. Cardenas[a], E. Castillo[b] and J. Meng[b]

[a]Fac. Ing. Eléctrica, Universidad Central de Las Villas, Carr. Camajuani KM 5 1/2, 50100 Santa Clara, Cuba

[b]University of New Brunswick, Room C119 Head Hall, 15 Dineen Drive, Fredericton, Canada E3B 5A3

julian@uclv.edu.cu

Power spectral subtraction (PSS) is a technique used to improve the signal-to-noise ratio in many applications in the recent years. This paper reports a successful application of a geometric-based PSS algorithm to improve the performance of automatic speech recognition under noisy conditions. The selected algorithm performs significantly better than other traditional spectral subtraction algorithms in the presence of low SNRs with low computational cost for speech enhancement. The performance contribution of the algorithm was assessed with CMU SPHINX-III speech recognition system using TIDIGITS speech corpus. Data was corrupted with seven noise types taken from the NOIZEUS database under seven different noise conditions (SNRs from -5 dB to 20 dB) for clean and multi-condition training setups. After extensive testing, results demonstrate that the selected algorithm is capable of improving recognition performance by 15% over the baseline approach at 0dB SNR when multi-condition training is used. The algorithm is particularly robust for noisy environments with low SNRs such as those present in car and airports. The algorithm is suitable to enhance the performance of speech recognition systems in mobile applications.

1 Introduction

Automatic speech recognition (ASR) has received considerable attention and has achieved outstanding performance in noise-free environments. However, under more realistic conditions where background, additive and convolutional noise is present; performance degrades significantly, discouraging its practical use. The literature on robust ASR discusses various approaches to cope with this problem. Some approaches attain robustness using one, or a combination of techniques that can be grouped as speech enhancement/preprocessing techniques, robust feature extraction methods, feature post processing techniques, and model adaptation to noisy environments.

Spectral subtraction (SS) methods belong to the class of speech enhancement techniques that have been largely applied in ASR contexts [1,2]. However, most speech enhancement methods are tailored at improving speech intelligibility for human listeners and hence, they may not perform very well in ASR tasks. These methods aim at improving the quality of the noisy speech by reducing the noise while minimizing the speech distortion introduced during the enhancement process. There are basically three sources of errors when spectral subtraction is applied to noisy speech, namely magnitude, cross-term and phase errors. Cross-term and phase errors are often neglected yielding to considerable degradations of recognition performance at low SNRs [2].

Loizou et al. [3] developed a new deterministic, geometric-based approach (GA) to spectral subtraction that addresses two fundamental shortcomings of the enhancement technique: the musical noise and invalid assumptions about the cross terms being zero. This paper investigates the effect of the application of a geometric-based power spectral subtraction method to enhance speech recognition accuracy under severe noisy conditions and demonstrates the effectiveness of the method at very low SNRs.

2 Methods

Figure 1 shows a block diagram of the ASR system used to assess the effect the power spectral subtraction algorithm on recognition performance. The experiments were designed to evaluate the benefits of combining the spectral subtraction method with feature post processing and multi-condition training to enhance the performance of ASR systems for mobile devices. The combined approach is contrasted with separate use of the techniques.

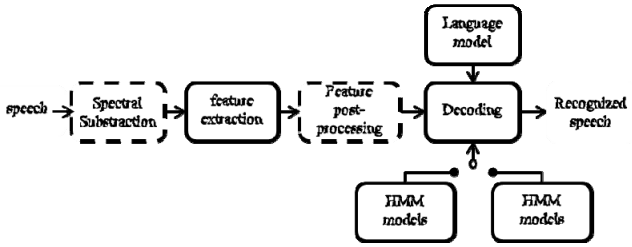

Figure 1: Block diagram of the ASR system used.

2.1 Spectral subtraction

The spectral subtraction algorithm selected for our experiments is the geometric-based power spectral subtraction algorithm developed by Lu & Loizou [3]. Objective and subjective evaluation of the algorithm using the NOIZEUS database [N] showed that the GA SS algorithm outperforms other traditional SS algorithms especially at low SNRs [3]. Although the primary application of the selected algorithm is speech quality improvement for human listeners, it has several advantages that make it attractive for other applications such as the preprocessing stage of ASR systems.

In [2] authors demonstrated that commonly neglected sources of error in SS algorithms can significantly affect recognition performance at SNRs around 0 dB. Lu & Loizou [1] also demonstrated that as the SNR$\rightarrow\pm\infty$, the effect of cross terms tends to vanish asymptotically; but they can be particularly large at SNRs where most speech enhancement algorithms operate. The GA power spectral subtraction algorithm [L] has the advantage of incorporating the cross terms involving phase differences between the noise and the noisy signals. Furthermore, the algorithm minimizes the musical noise typically present with other approaches. In addition, the selected algorithm is computationally efficient requiring few multiply and add operations. This make it suitable as a preprocessing stage of the automatic recognition systems used for mobile devices.

2.2 Speech corpus

The ASR experiments reported in this paper were performed on a speech corpus that used speech material from the TIDIGITS database and eight real-world noises from the NOIZEUS database. The TIDIGITS database contains over eight hours of high quality recordings of digit sequences spoken by 111 male and 114 female US-American adults. We used all recordings from TIDIGITS except the children utterances following the original division for training and testing sets of the speech corpus, each containing approximately half of the speakers.

The TIDIGITS database was downsampled from the original rate (20 KHz) to 8 KHz. The speech files were initially filtered with the modified Intermediate Reference System (IRS) filters specified by the ITU-T P.862 [I] and combined with the NOIZEUS corpus [N]. The filtering process allowed us adding the noise extracted from the NOIZEUS database to the clean speech from TIDIGITS without affecting the spectrum of the speech signals.

The NOIZEUS database contains 30 sentences corrupted by eight different real-world noises at 0 dB, 5 dB, 10 dB, and 15 dB SNRs. The noise includes suburban train noise, multi-talker babble, car, exhibition hall, restaurant, street, airport, and train-station noise; all of them randomly selected from the AURORA database [A]. Noise extraction was carried out by subtracting the clean utterances in this database from the noisy ones.

The recordings from the TIDIGIT corpus were contaminated with different noise types at 6 different SNR levels (namely: 20 dB, 15 dB, 10 dB, 5 dB, 0 dB, and -5 dB). We computed the energy of each recording from the TIDIGITS database and added the noise at the selected SNR according to the signal energy level.

2.3 Speech decoding

We used the CMU's SPHINX-3 speech recognition system [C] to carry out our experiments. The speech decoder of this tool is based on Hidden Markov Models (HMM) trained from acoustic features. In order to extract acoustic features, signal waveforms were segmented into 25 ms frames at 10 ms frame rate. Pre-emphasis filtering and Hamming window were applied to each frame. Feature parameterization used 13 Mel-frequency cepstral coefficients (MFCC features) in addition to deltas and delta-delta coefficients which resulted in a 39-dimensional feature vector. Digit context-dependent triphones were modeled using 3 state HMMs, 250 tied states and eight Gaussian mixtures per state. We used trigram language model with weight and word insertion probability experimentally determined at 12 and 0.1 respectively.

2.4 Training and test sets

In order to assess the performance of the selected preprocessing technique based on Figure 1, we designed various training and testing scenarios:

- Training on clean speech only (baseline clean training)
- Training on enhanced clean speech using spectral subtraction as enhancement technique (SS clean training)
- Training on clean and noisy speech (multi-condition –MC– training)
- Training on enhanced clean and noisy speech (SS+MC training)
- All the above scenarios with cepstral mean normalization (CMN) as feature post-processing.

Training on clean data allows us modelling of speech without the distortion of any type of noise. It generates adequate baseline models for comparative purposes. The highest performance should be obtained using these models while testing on clean speech. However, a significant deterioration of recognition accuracy is expected if speech comes from a noisy environment, mainly due to the mismatch between the training and testing conditions. Training on clean speech also permits checking how well the speech enhancement algorithm resembles the clean speech spectrum of the noisy speech. The second training scenario also uses clean speech for training acoustic models but, the GA SS is applied to the training data to allow accounting for the distortions potentially included by the SS algorithm.

Two multi-condition training setups permit assessing the advantage of using the spectral subtraction algorithm when speech models include information about the noise types that potentially might corrupt speech data under real-world environments. Finally, a new set of models are created using the same training setups as before, but CMN is also included as a feature post-processing technique. We selected CMN as it can partially compensate for the shift of the mean of the probability distributions of the parameters representing the speech with additive noise [A].

Multi-condition training typically leads to the best performance when training and testing is done under the same noise environments. However, we should not expect the same performance under different noise conditions. For this reason, and for MC training, we chose to train acoustic models using only a subset of the noise types available and the SNRs under consideration. This approach would show us how the new trained models would perform under unseen noisy conditions. The noise types selected as part of the training set were: restaurant, airport, babble and train noise. The clean training set was replicated and contaminated with each of these selected noise types at four different SNRs (20, 15, 10 and 5 dB). Very low SNRs (0 and -5 dB) were excluded from the training set. The new multi-condition training set had 16 new subsets plus the original clean training set.

Testing sets were created in the same way, but in this case, all noise types and SNRs were included. Various testing sets were needed in order to assess the effect of pre-processing (SS) and post-processing (CMN) techniques when applied together or separately. Each test set is then comprised of 49 subsets (a clean subset plus 8*6 contaminated subsets)

3 Results and discussion

In this work, we present the recognition performance expressed as word error rates (WERs). Figure 3 shows the performance impact of the GA SS algorithm when clean speech is used for training. Two training scenarios are depicted in this figure: baseline clean training and SS clean training. We checked the performance changes originated by applying SS to the test set only and to both the test and train sets. Graph lines represent average WERs over all SNRs including clean speech (SNR=∞). As expected the baseline system leads to the worst performance with an average WER of 69.87% (presented here for comparative purposes). The average WER obtained when SS is applied to the testing set is 51.39% while the average WER is 51.92% when SS is applied also to the training set. These results show that the application of the selected SS algorithm improves recognition performance on all noise types. However, there is no benefit applying the selected

speech enhancement technique to the clean speech used for training. Only the speech contaminated with airport and multi-talker babble noise exhibit better results, probably due to the speech-like characteristics of these noise types.

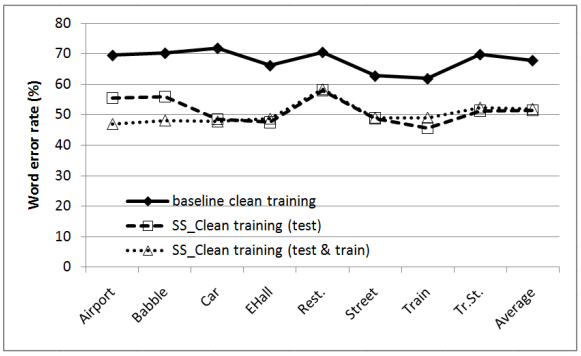

Figure 2: WER performance when training used clean speech

Figure 3 allows assessing the recognition impact of the combination of GA spectral subtraction with multi-condition training. In this case we continue using the clean training setup as the baseline performance. Except for the case of the MC training scenario, all the signals that comprise the test set were enhanced with the GA-SS algorithm. As aforementioned, the baseline system that uses clean speech for training achieved an average WER of 51.39%. In both MC training scenarios, performance behaves significantly better for all noise types. Note the slight WER increment experienced under acoustic conditions not included in the speech modeling (car, exhibition hall, street, and train station noise types). These results confirm the assumption that the system can perform well under unseen noisy environments.

The graphs in Figure 3 also evidence that the application of the spectral subtraction algorithm leads to improved recognition results. When only multi-condition training is used, the averaged WER was 23.99% which implies an improvement of 43.85% over the baseline system. The use of the geometric-based approach to spectral subtraction reduces WER in 48.44% (WER=19.42%) with respect to the same baseline system. Finally, if we compare performance improvement with respect to the results achieved when spectral subtraction was applied to the test set, and the baseline system was trained with the original clean speech; the impact of MC trainings amounts to 27.41% and 32% respectively.

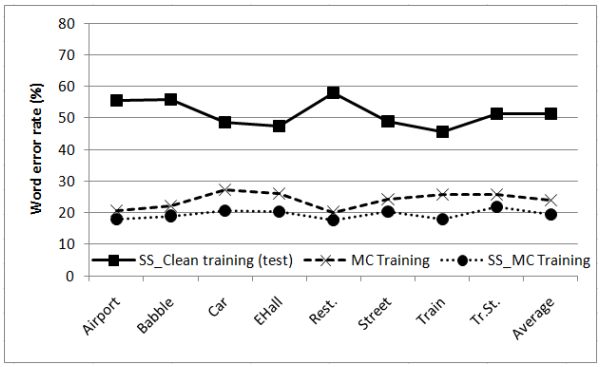

Figure 3: WER performance under different training conditions. No feature post-processing technique was used.

We also evaluated the effect of CMS as a feature post-processing technique. Figure 4 presents three graphs that show the effect of combining spectral subtraction, and CMN in conjunction with MC training. The use of CMN and clean training yielded an average WER of 46.57%. This result is better compared to the 51.39 % obtained under the same conditions but when SS was applied instead of CMN. Again airport and multi-talker babble environments were better enhanced. Both MC training scenarios slightly reduced the number of errors. Only a WER improvement of 0.93% was achieved after applying CMN in conjunction with GA SS and multicondition training.

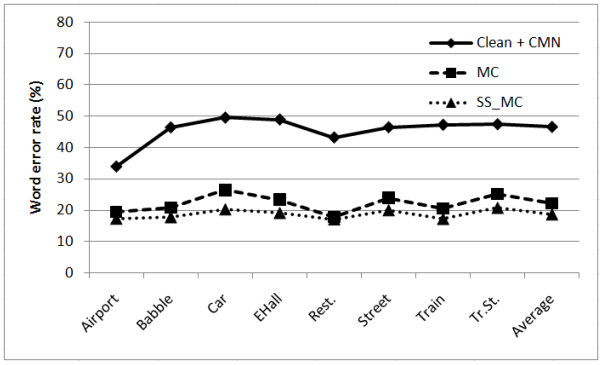

Figure 4: WER performance under different training conditions. CMN is applied as post-processing technique

Finally, we decided to evaluate the WER improvement of spectral subtraction over the baseline on MC training setups at different SNRs. Our results confirm the initial assumption that the geometric-based approach to power spectral subtraction enhances speech recognition accuracy at low SNRs. Figure 5 illustrates that the performance improvements can be up to 15% for low SNRs signals while it is small for high SNRs.

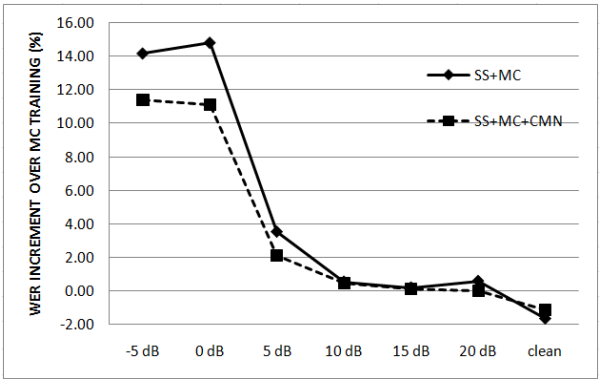

Figure 5: Averaged improvement on WER performance over all noise types at different SNRs.

5 Conclusions

We have presented experimental results that demonstrate the benefits of applying the geometric-based approach to power spectral subtraction in order to enhance the performance of speech recognition systems. Our results indicate that the combination of the selected power spectral subtraction algorithm, multi-condition training and feature enhancement with CMN increases the performance of the SRS under severe noisy conditions by up to 15%. The low computational cost of the selected algorithm makes it

attractive for mobile applications that use local speech recognition engines.

Acknowledgments

Authors would like to thank the Natural Sciences and Engineering Research Council of Canada (NSERC) and Version Two Inc. for the contributions made to this research.

References

[1] K. Paliwal, J. Lyons, S. So, A. Stark, K. Wójcicki, "Comparative evaluation of speech enhancement methods for robust automatic speech recognition," *4th International Conference on Signal Processing and Communication Systems*, (2010)

[2] D. Dimitriadis, J. C. Segura, L. Garcia, A. Potamianos, P. Maragos, V. Pitsikalis, "Advanced front-end for robust speech recognition in extremely adverse environments," *Proc. Interspeech- Eurospeech*,. 93, (2007).

[3] R. Gemello, F. Mana, R. De Mori, "Automatic speech recognition with a modified ephraim-malah rule," *IEEE Signal Processing Letters*, 13(1), 56–59, (2006).

[4] N. Evans, J. Mason, W. Liu, B. Fauve, "An assessment on the fundamental limitations of spectral subtraction", *In: Proc. IEEE Internat. Conf. on Acoustics, Speech, Signal Processing.* 1, 145-148 (2006)

[5] J. Flores, S. Young, "Continuous speech recognition in noise using spectral subtraction and hmm adaptation," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1, I/409–I/412 (1994).

[6] Y. Lu, P. Loizou, "A geometric approach to spectral subtraction", *Speech Communication,* 50, 453-466 (2008)

[7] Y. Hu, P., Loizou, "Subjective evaluation and comparison of speech enhancement algorithms". *Speech Communication*, 49, 588– 601, (2007)

[8] H. Hirsch, D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *Proc. 6th International Conference on Spoken Language Processing* (2000)