interspeech 2008 incorporating SST 08

Spectral Subtraction in Likelihood-Maximizing Framework for Robust Speech Recognition

Bagher BabaAli[1], *Hossein Sameti*[2], *Mehran Safayani*[2]

[1]Department of Computer Science and Engineering, Dashtestan Branch, Islamic Azad University
[2]Speech Processing Lab, Department of Computer Engineering, Sharif University of Technology
babaali@ce.sharif.edu, sameti@sharif.edu, safayani@sharif.edu

10.21437/Interspeech.2008-286

Abstract

Spectral Subtraction (SS), as a speech enhancement technique, originally designed for improving quality of speech signal judged by human listeners. it usually improve the quality and intelligibility of speech signals, while the speech recognition systems need compensation techniques capable of reducing the mismatch between the noisy speech features and the clean models. This paper proposes a novel approach for solving this problem by considering the SS and the speech recognizer as two interconnected components, sharing the common goal of improved speech recognition accuracy. The experimental evaluations on a real recorded database and the TIMIT database show that the proposed method can achieve significant improvement in recognition rate across a wide range of the signal to noise ratios.

Index Terms: Robust Speech Recognition, Signal Compensation, Spectral Subtraction.

1. Introduction

Automatic speech recognition has reached very high levels of performance in the controlled environmental conditions. However, the performance degrades significantly when environmental noise occurs during the recognition process. One approach to alleviate the performance degradation of speech recognition systems in noisy environments is prior enhancement of speech signal before being fed into the recognizer. Because the goal of these methods is both transforming the noisy signal to resemble clean speech and improving the quality of the speech signal, they called speech enhancement methods. We should use these methods as a pre-process unit in the speech recognition system. For instance spectral subtraction [1,2], Wiener filter [3], signal subspace [4], model-based speech enhancement [5,6] are widely used. Among the signal compensation methods, spectral subtraction is simple and easy to implement. Although it has low computational cost but it is very effective where the noise corrupting the signal is actually additive and varying slowly with time.

In conventional methods [7,8,9] of incorporating spectral subtraction as a signal compensation method in the front end of speech recognition systems, speech spectra are estimated by adjusting the spectral subtraction parameters according to the SNR. Generally, SNR is defined and computed over the whole input speech utterance. However, even if the noise is stationary, SNR varies according to speech energy. Therefore, the subtraction coefficient should be adjusted gradually according to the SNR. This method has been proposed as adaptive spectral subtraction [8]. Furthermore, in the SS, the subtraction factor is usually fixed for all the frequency bands under the assumption that the SNR is same throughout all frequency bands, although real world noise is mostly colored and does not affect the speech signal uniformly over the entire spectrum. Therefore, the subtraction factor should be adjusted according to the sub-band SNR. This method has been proposed as multi-band spectral subtraction [7]. In these approaches there is no feedback from recognition stage to the enhancement stage and they implicitly makes the assumption that generating a higher quality output waveform will necessarily result in improved recognition performance. However, the speech recognition is a classification problem, and the speech enhancement is a signal processing problem. So it is possible that by applying speech enhancement algorithms the hearing quality of the speech signal improves, but results in no improvement in recognition performance [10]. That is because the speech enhancement method may cause distortions in speech signal. For instance, in telephony speech recognition where speech models are trained by telephony speech that usually has low SNR, any signal compensation technique as judged by waveform level criteria will result in more mismatch between improved speech features and telephony speech model. Thus speech enhancement method improves speech recognition accuracy, only when it generates the sequence of feature vectors which maximize the likelihood of the correct transcription with respect to other hypothesis. So it seems logical that each improvement in pre-processing stage of the speech signal quality be according to the recognition criteria instead of waveform level criteria such as signal to noise ration or signal quality etc.

According to above reasons, in this paper a novel framework for applying multi-band spectral subtraction in front-end of the speech recognition systems is proposed. We show that by incorporating the speech recognition system into filter design process, the recognition rate is further improved. For implementing this scheme, we use an utterance the transcription of which is given and formulate the relation between spectral subtraction filter parameters and likelihood of the correct model. The proposed method has two phases, adaptation and decoding. In adaptation phase, spectral subtraction parameters is adjusted based on maximizing acoustic likelihood of the correct transcription and in decoding phase this optimized filter is applied for all incoming speech. Figure 1 shows the block diagram of the proposed framework.

The reminder of this paper is organized as follows. In section 2, we review the spectral subtraction technique. We derive the framework for maximum likelihood based spectral subtraction in section 3. Algorithm of proposed framework is described in section 4. Extensive experiments to verify the effectiveness of presented framework are presented in Section 5 and finally in section 6, we present summary of our work.

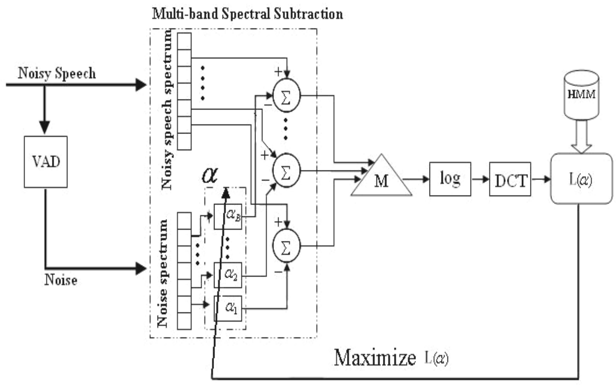

Figure 1: *Diagram of the proposed framework.*

2. Spectral Subtraction

Spectral subtraction is one of the most established and famous enhancement methods in removing additive and uncorrelated noise from the noisy speech. Spectral subtraction divides utterance to speech and nonspeech regions. It first estimates noise spectrum from nonspeech regions and then subtracts the estimated noise from noisy speech and produces the improved speech signal.
One of the most common methods used for estimating noise spectrum is [11]:

$$|N_n(k)|^2 = \begin{cases} (1-\lambda)|N_{n-1}(k)|^2 + \lambda|Y_n(k)|^2 & |Y_n(k)|^2 < \beta|N_n(k)|^2 \\ |N_{n-1}(k)|^2 & \text{Otherwise} \end{cases} \quad (1)$$

Where $|N_n(k)|$ is the absolute value of the kth coefficient of the Fourier transform of the nth noisy speech frame and $0 \le \lambda \le 1$ is the updating noise factor. β is the threshold parameter for determining the speech signal.
After estimating noise spectrum, we should estimate the clean speech spectrum as below:

$$|S_n(k)|^2 = \begin{cases} |Y_n(k)|^2 - \alpha|N_n(k)|^2 & \text{If } |Y_n(k)|^2 - \alpha|N_n(k)|^2 > \gamma|Y_n(k)|^2 \\ \gamma|Y_n(k)|^2 & \text{Otherwise} \end{cases} \quad (2)$$

where α is the spectral over-subtraction factor and is used to compensates the mistake in estimating noise spectrum. Therefore for getting better results we should set this parameter accurately and adaptively. γ is the spectral floor factor which is a small positive number guaranteeing that estimated spectrum will not be negative.
Since the noise signal does not affect the speech signal uniformly over the whole spectrum, it becomes imperative to estimate a suitable factor that will subtract just the necessary amount of the noise spectrum from each frequency band. Thus multi-band version of the spectral subtraction has been proposed which its operation is compatible by the operation of the human ear cochlea [12]. In this paper, we use Mel sub-band spectral subtraction that is reported to be more useful than full-band spectral subtraction [7].Whereas spectral over-subtraction factor is the most effective parameter in spectral subtraction algorithm, by adjusting this parameter for each Mel frequency band, we can expect remarkable improvement in performance of speech recognition systems.

3. Maximum likelihood based spectral subtraction (MLBSS)

Conventional spectral subtraction uses the waveform-level criteria such as maximizing signal to noise ratio, or minimizing mean square error and tries to decrease distance between noisy speech and desired speech. As we mentioned in the introduction, using these criteria should not necessarily decrease error rate. Therefore in this paper instead of waveform level criteria we use recognition error rate criteria for adjusting spectral over-subtraction parameters. One logical way for achieving to this goal is to select spectral over-subtraction vector so as to maximize acoustic likelihood of the correct hypothesis. This will increase the distance between the acoustic likelihood of the correct hypothesis and other competing hypothesis, so the probability that utterance be correctly recognized will be increased. Hence, in following the relation between spectral over-subtraction vector in pre-processing stage with acoustic likelihood of the correct hypothesis in decoding stage is formulated. These formulas depend on feature extraction algorithm and acoustic unit model. In this work, MFCC algorithm and hidden Markov model with mixture of Gaussian in each state are used for feature extraction and modeling of the acoustic unit respectively. Speech recognition systems based on the statistical model find the acoustic unit sequence most likely to generate observed feature vectors $Z = \{\mathbf{z}_1, ..., \mathbf{z}_t\}$ extracted from the improved speech signal. These observed features are a function of both incoming speech signal and also spectral over-subtraction vector. Speech recognizer gets the most likely hypothesis based on the optimal Bayes classification formula:

$$\hat{w} = \arg\max_{w} P(Z(\boldsymbol{\alpha})|w)P(w) \quad (3)$$

Where the observed feature vectors is a function of spectral over-subtraction vector $\boldsymbol{\alpha}$. In this formula $P(Z(\boldsymbol{\alpha})|w)$ and $P(w)$ are the corresponding acoustic and language score, respectively. Our goal is to find vector $\boldsymbol{\alpha}$ achieving the best recognition performance. Similar to either speaker or environmental adaptation methods, for adjusting $\boldsymbol{\alpha}$, we need some adaptation data with a known transcription. We assume that the correct transcription of the utterance w_C is known. So the value of $P(w_C)$ can be ignored because this value is constant regardless of the value of $\boldsymbol{\alpha}$. We can then maximize equation (3) with respect to $\boldsymbol{\alpha}$ as:

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} (P(Z(\boldsymbol{\alpha})|w_C)) \quad (4)$$

In a HMM based speech recognition the acoustic likelihood is estimated by single most likely state sequence. If S_C represents all state sequences in the combinational HMM and S represents single most likely state sequence, then the Maximum likelihood estimation of $\boldsymbol{\alpha}$ is written as:

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}, s \in S_C}{\operatorname{argmax}} \left\{ \sum_i \log(P(z_i(\boldsymbol{\alpha})|s_i)) + \sum_i \log(P(s_i|s_{i-1}, w_C)) \right\} \quad (5)$$

Regarding to (5) for getting $\hat{\boldsymbol{\alpha}}$ acoustic likelihood of the correct transcription should be jointly maximize with respect to the state sequence and $\boldsymbol{\alpha}$ parameters. This joint optimization should be performed in an iterative manner.

Noisy speech is passed to the spectral subtraction filter and feature vectors $Z(\boldsymbol{\alpha})$ is extracted given the known $\boldsymbol{\alpha}$. Then optimal state sequence $s = \{s_1, ..., s_t\}$ is computed using (5), given the correct transcription w_C. State sequence $\hat{s}$ simply can be computed using Viterbi algorithm. Given the known state sequence $\hat{s}$ we want to find $\hat{\boldsymbol{\alpha}}$ as:

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} \left\{ \sum_i \log(P(z_i(\boldsymbol{\alpha}) | \hat{s}_i)) \right\} \quad (6)$$

There is not a closed-form solution for computing optimal $\boldsymbol{\alpha}$ for a given state sequence, so we use non-linear optimization.

4. MLBSS Algorithm

In section 3 we present a novel approach for adjusting spectral over-subtraction vector so as to maximize acoustic likelihood of the correct transcription. Here we should answer a question. If the correct transcription was known a priori, there would be no need for recognition. For answering this question we should mention that the correct transcription is only needed in the adaptation phase and in the decoding phase these parameters are fixed. At first user says an utterance with a priori known transcription then the utterance is passed through the spectral subtraction filter fixed with initial parameters. After that most likely state sequence is generated using Viterbi algorithm. Then the optimum spectral subtraction filter is produced given the state sequence. Recognition is performed on validation set with using this optimized filter if the desired error rate is satisfied, the algorithm is finished otherwise the new state sequence is estimated. In an iterative manner, the spectral over-subtraction vector which maximizes the total log likelihood of the utterance with a known transcription is found. Feature vector at first is extracted from the improved speech signal and then the log likelihood is computed given the known state sequence. If the likelihood does not converge, the gradient of the spectral over-subtraction vector is computed, and the spectral over-subtraction vector is updated. Spectral subtraction is performed with using this updated vector and the new feature vectors are extracted. This process repeats until the convergence condition is satisfied.

In proposed algorithm like speaker and environment adaptation techniques, spectral over-subtraction vector adaptation can be implemented either in a separate off-line session or by embedding an incremental on-line step to the normal recognition mode of system. In off-line adaptation as explained in above, the user is aware of adaptation, typically by performing a special adaptation session, while in on-line adaptation the user may not even know that adaptation is carried out.

5. Experiments

In this section, proposed framework is evaluated and is also compared to the traditional spectral subtraction method for speech recognition. Speech recognition experiments conducted on the Nevisa [13], a large vocabulary, speaker-independent, continuous HMM-based speech recognition system has been developed in speech processing lab of computer engineering department of Sharif University of technology. Experiments have been done in two different operational modes of Nevisa system: phoneme recognition on TIMIT database and isolated command recognition on a distant talking database recorded in real noisy environment. The reason for performing phoneme recognition instead of word recognition is that in the former case, the recognition performance lies primarily on the acoustic model. The phoneme recognition accuracy is calculated as below :

$$Accuracy(\%) = \frac{N - S - D - I}{N} * 100\% \quad (7)$$

with S substitution errors, D deletions and I insertions for N test phonemes.

5.1. Evaluation on added noise condition

The first experiments are conducted on the TIMIT database. The test sets are obtained by the artificially adding of seven types of noises (Alarm, Brown, Multi-Talk, Pink, Restaurant, Volvo and White noise) from the NOISEX-92 [14] database to the TIMIT speech database. For evaluating our algorithm, its results are compared with Kamath's [12] spectral subtraction which is a SNR based method. From the experimental results, we observed the following facts. With regards to the difference between noise type and SNR, the result shows that the MLBSS method was capable to improve recognition performance compared to Kamath's method. In some cases, Kamath's method achieves lower performance than the baseline. This is due to spectrum distortions caused by not adjusting the spectral subtraction factors and destroy the discriminability in pattern recognition. The mismatch reduces the effectiveness of the clean trained acoustical models and causes the recognition accuracy to fall. Higher SNR difference between the training and testing speech causes higher degree of mismatch, and therefore result in greater degradation in recognition performance.

Table 1. *Phoneme recognition accuracy rate(%) on 100 sentences from the test set of TIMIT*

Noise	Method	0dB	5dB	10dB	15dB	20dB
Alarm	No Enhance	19.5	29.0	37.1	46.1	52.9
	Multiband SS	20.2	27.5	38.2	45.1	49.3
	MLBSS	21.4	32.0	40.4	48.3	55.0
Brown	No Enhance	40.6	48.5	56.5	61.9	64.0
	Multiband SS	40.7	49.1	53.3	54.7	55.0
	MLBSS	43.1	54.3	60.6	64.6	65.1
MultiTalk	No Enhance	14.0	23.6	34.4	43.8	51.9
	Multiband SS	14.0	25.8	35.2	43.0	49.5
	MLBSS	16.0	26.2	37.4	47.2	54.5
Pink	No Enhance	8.9	13.3	20.7	28.5	37.9
	Multiband SS	10.6	15.9	22.6	33.4	41.9
	MLBSS	11.2	17.2	24.8	33.8	43.1
Restauran t	No Enhance	13.3	20.8	30.2	40.0	48.9
	Multiband SS	16.3	25.5	33.8	40.3	46.1
	MLBSS	16.7	25.8	34.4	42.2	50.1
Volvo	No Enhance	44.4	48.6	53.4	57.4	61.0
	Multiband SS	43.0	48.3	51.1	52.9	54.2
	MLBSS	46.3	51.5	55.5	59.4	63.0
White	No Enhance	3.7	8.5	15.7	24.0	31.8
	Multiband SS	5.1	9.6	20.9	25.7	36.0
	MLBSS	5.9	11.9	22.4	28.3	36.3

5.2. Experimental results in clean environment

The performance of noise robust front-end processing is sometimes degraded under clean test conditions because speech enhancement methods such as spectral subtraction sometimes yield unexpected distortions for clean speech. So, even though the performance of an MLBSS algorithm is considerably good under noisy environments, it is not desirable if the recognition rate decreases for the clean speech. For this reason, we evaluated the performance of MLBSS algorithm not only in the noisy conditions but also when the speech was free from any noise on TIMIT. Recognition results obtained from the clean condition are shown in Table 2 where we can find that the recognition accuracy of the MLBSS approach is even a bit higher than that of the baseline while multi-band SS shows noticeable decrease. This phenomenon can be understood such that the MLBSS approach has the ability to compensate for the effects of noise, so only the mismatch be reduced.

Table 2. *Phoneme recognition accuracy rate(%) on 100 clean sentences from the testset of TIMIT*

No Enhance	Multi Band SS	MLBSS
66.43	63.75	66.79

5.3. Evaluation on data in real environment

It can be concluded from the above experiments that MLBSS algorithm has the capability to significantly increase the robustness of the recognition system on artificially noise added data. However, a direct comparison is still missing as the experimental setup is real environments. Therefore, a third set of experiments was performed and will be described in below.

We use Nevisa system in isolated command recognition mode trained with clean isolated command and tested with the noisy data captured from a microphone placed 2m distance from the speaker. We collect training dataset using a close-talking microphone in a quiet office by 16 females and 32 males who uttered 30 commands such as turning on/off or opening/closing different devices in an office. 10 males and 5 females who are different from training speakers, utter the commands at 2 m distance from centre of the microphone. Room dimension is (4.5m x 3.5m x 3.5m), characterized by an amount of reverberation ($T_{60} \cong 0.3s$). There are some sources of noise such as 3 computers and load speaker which is propagating office noise from NOISEX-92 database on 40 degree from the microphone. Average SNR of the dataset is 15db. Each command is modeled by fifteen states left-to-right HMMs with no skip (2 Gaussians/state). CMN is performed on the training utterance. Different results and experiments are shown in figure 2. From this experiment, we can make following observations. 1) Each approach is able to improve the robustness of the system. 2) MLBSS combined by CMN yields the highest robustness with respect to noise among the approaches investigated. 3) While the robustness of the MLBSS approach is slightly inferior to that of the SS, it yields better performance when combined by CMN.

6. Conclusions

A novel likelihood maximizing based spectral subtraction framework for robust speech recognition was proposed and verified. Experiments showed that the proposed approach could effectively improve recognition performance in noisy conditions, while maintaining the same or even yielding slightly better recognition accuracies in clean speech environment. The combination of noise removal approaches and cepstral mean normalization led to further improvement of recognition performance.

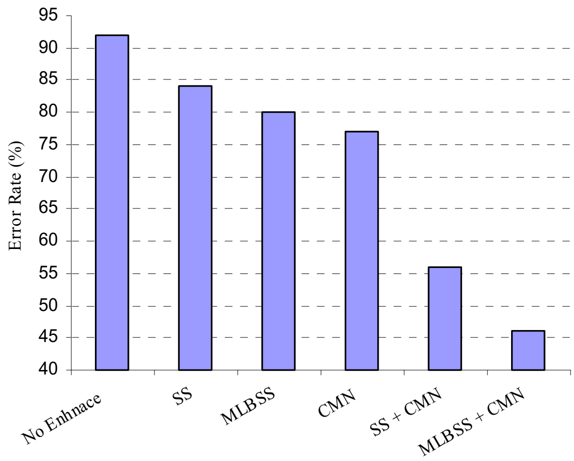

Figure 2: *Error Rate(%) in real environment with respect to different combinations of the MLBSS, SS and CMN. (SS stand for Kamath's multi-band spectral subtraction)*

7. References

[1] Boll, S.F., "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", IEEE Trans. on Acoustics, Speech and Signal Process. 27(2):113-120, 1979.

[2] Fujimoto, M., Ogata, J., Ariki, Y., "Large Vocabulary Continuous Speech Recognition under Real Environments Using Adaptive Sub-Band Spectral Subtraction", in ICSLP, Vol.I, pp.305-308, 2000.

[3] Stahl, V., Fischer, A., Bippus, R., "Quantile based noise estimation for spectral subtraction and Wiener filtering", in ICASSP, vol. 3, pp. 1875-1878, 2000.

[4] Hermu, K., Wambacq, P., Van hamme, H., "A review of signal subspace speech enhancement and its application to noise robust speech recognition", EURASIP Journal on Applied Signal Processing archive, Volume 2007 ,Issue 1, 2007.

[5] Sameti, H., Sheikhzadeh, H., Deng, L., Brennan, R. L., "HMM-Based Strategies for Enhancement of Speech Embedded in Non-Stationary Noise", IEEE Trans. on Speech and Audio Proc., 6(5):445-455, 1998.

[6] Stouten, V., Van hamme, H., Wambacq, P., "Model-based feature enhancement with uncertainty decoding for noise robust ASR", Speech Communication Journal. 2006.

[7] Chen, J., Paliwal, K. K., Nakamura, S., "Sub-Band Based Additive Noise Removal for Robust Speech Recognition", in Eurospeech, 2001.

[8] Yamamoto, H., Yamada, M., Komiri, Y., Ohora, Y., "Estimated Segmental SNR base Adaptive Spectral Subtraction Approach for Speech Recognition", Technical Report of IEICE in Japan, SP94-50, pp.17-24, 1994.

[9] Kleinschmidt, M., Tchorz, J., Kollmeier, B., "Combining speech enhancement and auditory feature extraction for robust speech recognition", Speech Communication 34, 75-91, 2001.

[10] Seltzer, M., Raj, B., Stern, R.M., "Likelihood-Maximizing Beamforming for Robust Hands-Free Speech Recognition", IEEE Trans. Speech and Audio Proc., 12(5):489-498, 2004.

[11] Berouti, M., Schwartz, R., Makhoul, J., "Enhancement of speech corrupted by acoustic noise", Proc. IEEE Int. Conf. Acoust, Speech, Signal Process., pp. 208-211, Apr. 1979.

[12] Kamath, S., Loizou, P., "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise", in ICASSP, 2002.

[13] Babaali, B., Sameti, H., "The Sharif Speaker Independent Large Vocabulary Speech Recognition System", The 2nd Workshop on IT & Its Disciplines, Kish Island, Iran, 2004.

[14] Varga, A., Steenneken, H.J.M. , Tomlinson, M., Jones, D., “The Noise-92 Study on the Effect of Additive Noise on Automatic Speech Recognition,” DRA Speech Research Unit, St. Andrew’s Rd., Malvern, Worcestershire, WR14 3PS UK,1992.