Adaptive Speech Enhancement with Varying Noise Backgrounds

Thorsten Kühnapfel 1, Tele Tan 1, Svetha Venkatesh 1, Sven Erik Nordholm 2, Burkhard Igel 3

1 Dept. of Computing, Curtin University of Technology, Western Australia
2 Western Australian Telecommunications Research Institute, Western Australia
3 Dept. of Information Tech. and Electrical Eng., University of Applied Sciences Dortmund, Germany

Abstract

We present a new approach for speech enhancement in the presence of non-stationary and rapidly changing background noise. A distributed microphone system is used to capture the acoustic characteristics of the environment. The input of each microphone is then classified either as speech or one of the predetermined noise types. Further enhancement of speech in respective microphones is carried out using a modified spectral subtraction algorithm that incorporates multiple noise models to quickly adapt to rapid background noise changes. Tests on real world speech captured under diverse conditions demonstrate the effectiveness of this method.

1 Introduction

Audio sensors provide an effective and low cost way of measuring the sonic activity in and around places of interest. Whether used on its own, or with other sensors (CCTV, motion sensors), the quality of the sound captured in the presence of interfering noise signals is an important issue. To compensate for ambient noise, multiple microphones can be used to leverage on the detection and localisation of the speaker or speakers. One technique is to use beam steering or blind source separation to isolate the source of the speaker/s. However, the main drawback of this approach is the inability to work with sources that are not close to the microphone. One may consider distributing the microphone arrays in the area of interest, but this is costly. This paper explores the use of an alternative method of using several microphones to capture the noise and use the closest microphone to enhance speech signals.

For single microphone speech enhancement, spectral subtraction filters are commonly used, in which a single estimation of the background noise is made. There are different techniques to model the background; heuristically [5, 2, 9, 10, 6] by averaging out the recorded noise sequences or statistically by modelling each background coefficient as Gaussian random variables [4]. Recent work has focused on improving the filter coefficients [2] or investigating various smoothing techniques [9] to minimise speech distortion, known as "musical noise". The authors in [10] use a modified spectral subtraction approach with a low resolution gain function which is smoothed over time. The filter proposed in [6] divides the signal in different frequency bands and uses a weighting function to adjust the subtraction factor for each sub band. The limitation of using only one background model is that in a real world situation, the noise source can change rapidly and all the above methods require time to adapt.

To successfully use a spectral subtraction algorithm, voice activity detection (VAD) is essential. Martin [7] has reported a fast and effective algorithm for estimating the SNR based on short time power estimation. A disadvantage of this approach is that noise intensity estimation is sensitive to outliers and can lead to false detections. A more sophisticated VAD method is proposed in [8]. It uses a statistical approach to compare the second-order statistics of the test signal to the speech model. This is, however, complex and computationally expensive and has only been tested on stationary artificial or slightly non-stationary helicopter noise.

The aim of the proposed system is to provide sonic surveillance for an area of interest. As the size of this area increases it will be more economical to use a network of distributed microphones, instead of several microphone arrays. Within this network, the closest microphone to the source is used to enhance the speech signal, whereas the other microphones are used to capture and classify the noise source. The system first performs speech/non-speech classification using a new voice activation detection algorithm on all microphone inputs independently. For the non-speech segments, further classification into a predetermined list of background models is carried out. This classification result is used to provide the appropriate noise model to enhance the quality of the classified speech segments. The VAD is also used to set the parameters of the spectral filter. The experiments show the reliability of the noise classification and speech detection, in presence of real, non stationary background noise. The enhanced speech is then evaluated by a group of 11 people.

The novelty of the proposed system is that different noise models are constructed for the spectral subtraction algorithm, and thus the system rapidly adapts...
to changes in typical noise sources in the environment. This is in contrast to other systems that need time to adapt to noise changes. Therefore, it can be used to increase the performance of any current spectral subtraction algorithms. Additionally, the noise classification is incorporated within the VAD to minimise false detection of speech.

2 Methodology

The proposed system consists of three sub modules: Noise classification, voice activity detection and speech enhancement. The noise classification is done by matching the features of the noise models with the input signal $y$. This correlation value is also used to enhance the reliability of the voice activity detection (see figure 1). Speech enhancement is then performed by combining the noise classification result of several microphones in conjunction with voice activity detection.

2.1 Noise Classification

For noise classification, the recorded signal $y_k(i)$ is transformed into the frequency domain $Y_k(i,f)$ via the fast Fourier transformation, where $i$ is the time block index, $f$ the frequency index and $k$ is the index of the microphone. $|Y_k(i,f)|$ is then scaled by a Mel scale triangular filter [3] to obtain the final feature set $S_k(i,p)$, where $p$ is the Mel scale filter index. In the initial learning process, a mean noise model $N_k$ for microphone $k$ and noise type $q (q: 1...Q)$ is computed as

$$N_k(i=0,p) = \frac{1}{\tau+1} \sum_{j=1}^{\tau} S_k(j,p)$$

where $\tau$ is the numbers of time blocks used. The classification decision $\eta_k$ for microphone $k$ is made by computing a normalised cross correlation coefficient $c_k^q$ between each noise model $N_k$ and the current feature set $S_k$ and is computed as:

$$\eta_k(i) = \text{argmax}_q \{c_k^q(i)\}$$

If the correlation coefficient $c_k^q$ of the classified noise $\eta_k$ is above a threshold (> 0.95), it indicates a single noise source and the corresponding noise model is updated as

$$N_k(i,p) = (1 - \rho)N_k(i-1,p) + \rho S_k(i,p)$$

where $\rho$ is an exponential updating factor.

Let $l^q$ be the count of noise type $q$ across the $K$ microphones in which no active speech is detected. Then the final classification of the overall noise type is based on $\text{argmax}_q \{l^q\}$.

2.2 Voice Activity Detection

The final VAD is based on two features, the signal power, $P_y$, and the correlation coefficient $c_k^q$ of the detected noise $\eta_k$ of microphone $k$. Figure 1 shows the generic representation of the voice activity.

2.3 Spectral Subtraction

The recorded signal $y(n)$ can be considered as the sum of the speech signal $x(n)$ and uncorrelated noise signal $\omega(n)$ as

$$y(n) = x(n) + \omega(n)$$

where $n$ is the sample. To estimate the speech-only signal $\tilde{x}$, a time varying filter with gain function $G$ can be applied to the short-term frequency spectra $Y$ as

$$\tilde{x}(i,f) = G(i,f)Y(i,f)$$
where \( i \) is the time block index and \( f \) the frequency bin. 
\( G \) is defined by the magnitude spectra of the modelled noise \( P_{\nu}(i, f) \) and the input signal \( P_{Y}(i, f) \) [5] as

\[
G(i, f) = \sqrt{1 - \frac{P_{\nu}(i, f)}{P_{Y}(i, f)}}
\]

(12)

A disadvantage of such a filter is that it introduces a distortion in the speech signal known as residual noise or “musical noise”. To smooth out the level of the residual noise, a subtraction factor \( \alpha \) and a spectral floor function \( \beta \) is added to the gain function [1] as

\[
G(i, f) = \max \left\{ \sqrt{1 - \frac{P_{\nu}(i, f)}{P_{Y}(i, f)}} \beta \right\}
\]

(13)

To enhance the accuracy of the filter, the modelled noise spectra is estimated over time [5].

### 2.4 Speech Enhancement

The speech enhancement is done by background noise removal via spectral subtraction. In our case, for each noise type \( q \), a modelled background noise \( P_{k}^{q} \) of the magnitude spectra is estimated. These different models enable the system to adapt quickly to changes in environmental noise. Also the VAD is used to define three different sets of parameters for the gain function to enhance the filter performance.

\[
G_k(i, f) = \max \left\{ \sqrt{1 - \frac{\eta q(i, f)}{\nu_k(i, f)}} \beta_k \right\}
\]

(14)

where

\[
\alpha = \begin{cases} 
\max \sqrt{1 - \frac{\eta q(i, f)}{\nu_k(i, f)}} & \text{if } \nu_k(i) = \psi_{no} \\
\max \sqrt{1 - \frac{\eta q(i, f)}{\nu_k(i, f)}} & \text{if } \nu_k(i) = \psi_{high} \\
\max \sqrt{1 - \frac{\eta q(i, f)}{\nu_k(i, f)}} & \text{if } \nu_k(i) = \psi_{low}
\end{cases}
\]

The parameters of \( G \) for sequences with no speech, \((\nu_k = \psi_{no})\) are stricter to remove noise components by about 10 to 20 dB. Also, the adaptation rate of the modelled noise \( P_{k}^{q} \) can be set quite high because the signal contains only noise and no other signals. In cases where speech is detected, the gain function is less strict to preserve the speech signal and to minimise the residual noise. The adaptation rate of the noise model is set to 0 for \( \psi_{high} \) or to a low value for \( \psi_{low} \).

### 3 Experiments

The performance was evaluated in four different experiments on real world data and one experiment on synthetic data. Audio was recorded at 16 kHz and 16 bit/s. All experiments on real world data used the same data set: 5 distributed microphones in a 5.8m x 7.3m room. The normal background noise consists of air-conditioning and PC-ventilation noise. Additionally, recorded noise in a cafe during lunch time and scooter noise was played back into the environment. During the recording, two persons were asked to talk near different microphones, we call them speaker 1 and 2.

#### 3.1 Noise Classification

The common noise pattern for all experiments is cafe noise from the beginning to 41.1s, scooter noise till 50.3s, then cafe noise again to about 57s and finally scooter noise to the end, which also fades in and out. Further, the noise level increases between 34.3s and 38.3s and between 45.3s and 48.7s.

![Figure 2. Cafe noise, light green (1); scooter noise, dark blue (2). Y-Axis shows certainty of \( \eta q \).](image)

Figure 2 shows the classification result over all available microphones where no speech is detected. At about 15.5s the noise classification result varies between the individual microphones. This is shown as a small peak.

#### 3.2 Voice Activity Detection

This experiment demonstrates the voice activity detection algorithm for two speakers standing at different microphones, and the ability to handle rapid changes in noise levels.

![Figure 3. Dark red (1) marks speech sequences, active speech labelled as 0.1. Light blue (2) indicates speech with high energy. Dark blue (3) is the recorded signal and s indicates the ground truth of speech.](image)

Figure 3 shows the detected speech sequences for the closest microphones. Other microphones do not have any detected speech and are not shown. Even though cafe noise consists of multiple people talking, the false detection rate was 6% and 3% for speaker 1 and 2 respectively.
3.3 Synthetic Changing Noises

This experiment demonstrates the advantage of multiple noise models. Three different noise sources, band limited white noise, are synthetically mixed with a clean speech signal. Only one noise source is present at any time and noise changes occur during speech sequences.

Figure 4. a) SNR of the mixed signal (1) and the SNR of the clean speech signal (2) to the introduced noise. b) SNR of the estimated speech signal with 1 noise model (1), 2 noise models (2) and 3 noise models (3). Frequency range of the white noise is 300 to 800 Hz (n1), 2.5 to 3 kHz (n2) and 5 to 5.5 kHz (n3).

Figure 4b shows that the estimated speech signal is best when the number of noise models match the number of noise sources. We measure the variance in SNR of the estimated speech signal to the clean speech signal (2) of figure 4a in speech sequences with changing noise. The result was a variance of 7.41 dB, 4.82 dB and 2.05 dB for 1, 2 and 3 modelled noise sources respectively.

3.4 Changing Noise

To verify the results of experiment 3.3, this experiment used real noise sources and measured the signal energy of a noise-only sequence. In this sequence the noise type switches rapidly at about 41 seconds from cafe noise to scooter noise.

Figure 5. Dark green (1) is the noise energy, yellow (2) is the signal energy of the filtered noise signal with 1 noise model and in dark blue (3) with 2 noise models.

Figure 5 shows that it takes up to 1 second for the background model to adapt to the new noise source, wherein our system switches almost instantaneously. For sequences with speech this takes even longer because the noise model can not be updated during speech.

3.5 Qualitative Evaluation

The final noise suppressed output of the system is evaluated by a group of 11 people. They listened to the original and the enhanced signal, and gave a mean opinion score (MOS), expressed as a single number from 1 (Bad) to 5 (Excellent). The enhanced signal achieves an average score of 4.3 and the original recording an average score of 2.2. A video file that demonstrates the proposed system can be downloaded from: www.computing.edu.au/~thorsten/SpeechEnhancement.avi

4 Conclusion

We have demonstrated that the proposed system is able to detect and enhance speech sequences of a single microphone, in environments with non-stationary and rapidly changing background noise sources. The noise classification based on the entire network overcomes the problems of false detection. Our future work will investigate learning techniques to dynamically adapt to noise sources. Further, sophisticated VAD for low SNR situation will be considered to aid the elimination of heuristic thresholds.

References