SPEECH ENHANCEMENT USING A MULTI-MICROPHONE SUB-BAND ADAPTIVE GRIFFITHS-JIM NOISE CANCELLER

P. W. Shields and D. R. Campbell

Department of Electronic Engineering and Physics, University of Paisley,
High Street Paisley, Renfrewshire, PA1 2BE, SCOTLAND, U.K.
paul@diana22.paisley.ac.uk

ABSTRACT
The goal for most speech enhancement techniques is to improve speech quality by increasing intelligibility and/or reducing listener fatigue. The Multi-microphone Sub-band Adaptive (MMSBA) processing scheme has been shown to improve both the SNR and speech intelligibility of a wanted speech source in the presence of an unwanted noise source, in a moderately reverberant real-room environment using two microphones [1].

The proposed Sub-band Adaptive Griffiths-Jim (SBAGJ) processing scheme uses features from the Griffiths-Jim [2] adaptive noise canceller, in conjunction with the MMSBA scheme. The combination of the Griffiths-Jim front end, and the MMSBA noise cancelling section results in a system with the improved filter performance of a sub-band processing scheme, but bypasses the need to explicitly deal with the causality issues associated with many adaptive noise cancellation algorithms.

1. INTRODUCTION
There are many applications, such as, echo cancellation telecommunications, or hearing aids, where it is necessary to recover a wanted speech source from a signal corrupted an unwanted noise source. This problem is further complicated if there are unknown reverberation channels.

The Griffiths-Jim beamformer [2] has been applied to the task of improving speech quality in a noisy reverberant environment by Greenberg and Zurek [3] and Greenberg et al [4]. The Greenberg and Zurek work is a wide-band approach whereas in the latter publication the input signals are low and high pass filtered before processing. In this second approach, low frequency signals are binaurally processed, whereas the high frequencies are processed using an adaptive array. The result is equivalent to an increase of 3dB and 5dB respectively as measured using speech reception in noise when compared to binaural and monaural audition respectively.

In this investigation, SNR improvement scores for the proposed SBAGJ speech enhancement scheme are compared to the wide-band LMS (WBLMS) algorithm and the wide-band Griffiths-Jim (WBGJ) algorithm. All three techniques are applied to the task of improving the SNR of a wanted speech source in the presence of an unwanted noise source in a moderately reverberant real-room environment.

2. THE SBAGJ PROCESSING SCHEME

2.1. Real-Room Recordings
A real-room was chosen to make the speech and noise recordings. The room was as described in Figure 1, with soft furnishings present in the form of tables and chairs etc. These features were included to create as “typical” a living room situation as possible. The recordings were made using a KEMAR manikin, Knowles Model Number DB-4004, with Brul and Kjaer Type 4134 microphones fixed within the ear canal using a Zwislocki occluded ear simulator, Knowles Model Number DB-1001. Thus, a form of simulated head was present during the recordings. The reverberation time ($T_{60}$) for the real-room was established experimentally to be approximately 0.3s. The reverberation time chosen compares to the figure of: $T_{60}$=0.35s, which is representative of a typical domestic living room [5]

Figure1: Geometry for real-room recordings.
2.2. Sub-band Decomposition

Figure 2 illustrates the SBAGJ scheme with its sub-band decomposition process. The use of sub-bands reduces the problem of identifying a single long duration differential transfer function, to that of identifying a set of less complex parallel filters. This approach improves the convergence of the adaptive algorithm while controlling the computational cost. For reasons of experimental flexibility, the sub-band decomposition process was performed by splitting the two input channels into frames of length 256 samples, then applying a 50% overlap add FFT scheme. The frequency bins were then separated into 16 sub-bands with cochlear spacing prior to reconstructing the sub-band time domain signal. Using this approach, the error due to the process of decomposing the signal into frequency limited sub-bands and then synthesising a wide band signal was found to be in the order of 10^{-3} \%. The model used for the cochlear function was that of Greenwood [6]:

\[ F(x) = A(10^{ax} - k) \text{ Hz} \]  \hspace{1cm} (1)

\( F(x) \) is the upper and lower cut off frequency for each “cochlear” filter (A=165, a=2.1, k= 0.88.)

Figure 2: Sub-band Adaptive Griffiths-Jim Scheme.

2.3. Sub-band Processing Scheme

Figure 3 represents one of the Sub-Band Processing blocks (SBP) depicted in Figure 2. The inputs to each of the SBP blocks are generated from the sum and difference of the two microphone inputs, as described by Griffiths and Jim. Preparing the inputs to the system in this way avoids the inherent causality problem encountered with the LMS and similar algorithms.

The processing method employed in each SBP block depends on the cross-correlation/coherence between the channels. This allows the lower frequency bands which generally have high coherence (>0.7), to use an adapt and freeze strategy during a predetermined noise alone period (~ 0.4 seconds in this case), to adapt to the differential acoustic transfer function of the noise masker. The voice activity detector described by Agaiby [7] was used to determine the presence of a “noise alone” which was used for period, which was used for filter adaptation. The adaptive filter algorithm implemented was the LMS algorithm [8]. When speech is present, the weights in the adaptive filter are frozen, to allow preferential filtering out of the noise signal, leaving ideally only desired speech at the output E. The step-size \( \mu \) of the LMS algorithm was calculated for each individual band dependent on the variance of the band-limited reference channel input signal, \( \sigma_b \). The variance for each sub-band was calculated using the general recursive estimate:

\[ \sigma_b^2 = (1 - \alpha)\sigma_{b,1}^2 + \alpha \sigma_i^2 \] \hspace{1cm} (2)

where 0 < \( \alpha \) < 1 is a “forgetting factor” typically set at 0.95. In some of the higher frequency bands the speech information generally has a higher coherence than the noise source. This can take advantage of an approach described by Ferrara and Widrow [9]. In these bands the system is continually adapted to enhance the correlated component of the signal in each sub-band, which should emphasise the desired speech signal at the filter output \( Y \).

Figure 3: LMS Sub-band Processor (SBP)

Three example conditions and processing techniques are listed below:

1. Noise level below predetermined threshold; do not process sub-band, allow signal to pass unfiltered.
2. Noise signals highly correlated between channels; when no speech is present in either channel adapt the filter to minimise the noise signal then process with the frozen converged filter when speech is present.
3. Noise signals uncorrelated between channels; sum all bands of this type and continually adapt the filter, its output being used to estimate the speech signal.

In the experiments presented in this paper, the WBLMS and WBGJ schemes were processed according to condition 2. A combination of conditions 2 and 3 were used for processing using the SBAGJ scheme, the first 8 sub-bands were processed using the adapt and freeze strategy, with the remaining 7 bands summed together before processing using the Ferrara and Widrow approach.
3. TESTS AND RESULTS

Ten different sentences from the FAAF speech corpus [10] were recorded within the real-room environment described above. All signals were sampled at 20 kHz with 12-bit resolution. The masking noise source was speech shaped, generated by shaping the bins of a white noise signal to the long-term spectrum of the 80 FAAF sentences. The WBLMS, WBGJ, and SBAGJ algorithms were tested with the noisy reverberant speech signals at signal to noise ratios of –6dB, 0dB and +6dB. The filter length, W, of the adaptive filters was examined at lengths of 128, 512 and 1024 taps. For all signals, the SNR is measured as a long-term average across a standard selected section of the record. Where:

$$\text{SNR} = 10 \log_{10} \frac{\sigma_s^2}{\sigma_n^2}$$  \hspace{1cm} (3)

Analysis of the experiment aims to determine if there is a significant enhancement due to processing using the SBAGJ scheme, when compared to the unprocessed, WBLMS, and WBGJ schemes, when results are blocked by SNR. Secondly, to determine whether the adaptive filter length has a significant effect on filter performance.

Figures 4, 5, and 6 show the average SNR improvements using the SBAGJ and two wide-band schemes at –6dB, 0dB, and +6dB respectively. In each graph the SNR improvements are plotted using 128, 512 and 1024 taps respectively. In all cases, the error bars indicate the 95% confidence intervals.

At –6dB, 0dB and +6dB the wide-band schemes consistently score between approximately 1.5dB and 3dB SNR improvement. However, the SBAGJ scheme scores between 7.5dB and 8.2dB SNR improvement. This improvement is statistically significant at the 95% confidence level for all three SNRs.

In general all three schemes demonstrate a small average increase in SNR improvement when the adaptive filter length is increased from 128 through to 1024 taps. Table 1 shows the probability, p, of the difference in SNR improvement being statistically significant from a series of paired t-tests (two-tailed) comparing SNR improvement with adaptive filter lengths of 128, 512, and 1024 taps.

Both the WBLMS and WBGJ schemes show a statistically significant difference in SNR improvement when the adaptive filter length is modified. In six out of nine comparisons, the WBLMS demonstrates a difference in mean score at least at the 95% confidence level. The WBGJ scheme demonstrates a significant difference at least at the 95% confidence level in all nine cases. The SBAGJ processing scheme only demonstrates a statistically significant change in two comparisons out of nine. The ability of the SBAGJ
scheme to maintain the high level of performance using shorter filters, may be attributed to the system's ability to exploit the highly correlated speech in the higher frequency bands, processed using the Ferrara Widrow approach.

<table>
<thead>
<tr>
<th>SNR = -6dB</th>
<th>WBLMS</th>
<th>WBGJ</th>
<th>SBAGJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>128 vs. 512 taps</td>
<td>1.000</td>
<td>0.999</td>
<td>0.501</td>
</tr>
<tr>
<td>128 vs. 1024 taps</td>
<td>0.999</td>
<td>1.000</td>
<td>0.858</td>
</tr>
<tr>
<td>512 vs. 1024 taps</td>
<td>1.000</td>
<td>0.953</td>
<td>0.365</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SNR = 0dB</th>
<th>WBLMS</th>
<th>WBGJ</th>
<th>SBAGJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>128 vs. 512 taps</td>
<td>0.871</td>
<td>0.999</td>
<td>0.538</td>
</tr>
<tr>
<td>128 vs. 1024 taps</td>
<td>0.959</td>
<td>1.000</td>
<td>0.773</td>
</tr>
<tr>
<td>512 vs. 1024 taps</td>
<td>0.781</td>
<td>1.000</td>
<td>0.765</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SNR = +6dB</th>
<th>WBLMS</th>
<th>WBGJ</th>
<th>SBAGJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>128 vs. 512 taps</td>
<td>1.000</td>
<td>1.000</td>
<td>0.998</td>
</tr>
<tr>
<td>128 vs. 1024 taps</td>
<td>1.000</td>
<td>1.000</td>
<td>0.994</td>
</tr>
<tr>
<td>512 vs. 1024 taps</td>
<td>0.657</td>
<td>0.982</td>
<td>0.502</td>
</tr>
</tbody>
</table>

Table 1: Paired t-test comparison for WBLMS, WBGJ, and SBAGJ schemes using different filter lengths. Probability, p, of filter length providing a significant difference in SNR improvement.

4. CONCLUSIONS

The SBAGJ processing scheme has been shown to improve the SNR of a wanted speech signal corrupted with noise in a moderately reverberant environment by approximately 8dB. The SBAGJ provides a statistically significant performance gain over the WBLMS or WBGJ algorithms. It has been shown that using the acoustic scenario chosen for experimentation, where the speech source is in relatively close proximity to the microphones, that it is possible to maintain high levels of SNR improvement with a short adaptive filter using a diverse sub-band processing.

ACKNOWLEDGEMENTS

The authors thank for their advice and support: Mr. J Foster of the Institute of Hearing Research (IHR) at the University of Nottingham for assisting with the real-room recordings, and Professor S. Gatehouse of the IHR at Glasgow Royal Infirmary who supplied the speech material and subject contact. D.R. Campbell was supported by a Leverhulme Trust Fellowship during part of this work.

REFERENCES


