Abstract

Speech processing is at last starting the transition to wider bandwidths, whose benefits include increased intelligibility and comprehension, and a more pleasing communication experience. High quality, full-duplex Acoustic Echo Cancellation (AEC) is an integral component of a hands-free speakerphone communication system because it allows participants to converse in a natural manner, as they would in person. Some high-end telepresence systems already achieve life-like communication, but they are computationally demanding and prohibitively expensive. The challenge is to develop a robust Acoustic Echo Canceller (AEC) that processes full-band audio signals while maintaining low computational complexity and reasonable memory consumption for an affordable telepresence experience.¹

Introduction

High quality, full-duplex Acoustic Echo Cancellation (AEC) is an integral component of a hands-free speakerphone communication system, because it allows participants to converse in a natural manner, without a headset, as they would in person.

Echoes occur because there is an acoustic path between the loudspeaker and the microphone, as illustrated in a typical AEC system shown in Figure 1. The reference signal, \( x(n) \), is rendered by the loudspeaker in the near-side room. The microphone signal, \( y(n) \), contains the local talker’s speech signal, \( s(n) \), along with background noise, \( b(n) \) and the disturbances created by the unwanted echo, \( d(n) \). The AEC predicts the echo, \( \hat{d}(n) \), and subtracts it from the microphone signal to create \( e(n) \), which is typically further processed by residual echo suppression, noise reduction, and other speech enhancement algorithms.

Benefits of Wider Bandwidth

The benefits of using wider bandwidth for speech communication are described well by Pennock and Hetherington², and in the works they cite. The key improvements of wideband over traditional narrowband are that wideband speech can increase intelligibility and comprehension, and create a better sense of presence by sounding more like in-person, face-to-face communication. High frequency

¹ This paper was first presented at the AES 128th Convention, London, 22-25 May 2010.

content is preserved, which can reduce the incidence of confusion, especially between consonants.

When referring to speech, the term “High Definition” (or “HD”) has been used by many people in many different ways, loosely meaning anything more than narrowband speech. To avoid confusion, we define the following terms for speech signals and their respective bandwidths: narrowband, 300-3,400 Hz; wideband, 50-7,000 Hz; super-wideband, 50-14,000 Hz; full-band, 20-20,000 Hz.

**Technical Challenges**

Our existing Aviage Acoustic Processing software library (released in summer 2009) already processes wideband speech operating at a 16,000 Hz sample rate. Our challenge was to upgrade the library so that it would handle full-band speech operating at a 48,000 Hz sample rate, and achieve this on existing systems without increasing the computational complexity or consuming more memory (i.e., maintain the same MIPs and memory usage). The underlying audio path including the AD/DA converters, microphones, amplifiers, and loudspeakers was already designed to handle full-band signals.

**Brief Survey Of Solutions**

Beaugeant et al. describe a “dual-mode” solution that splits a wideband signal into low band (0-4,000 Hz) and high band (4,000-8,000 Hz) regions. This solution offers the advantage of backwards compatibility with narrowband systems. Unfortunately, however, it does not address an extension to a full-band solution, and it does not describe the methods it uses to stay within the same computational and memory limits of a narrowband system, albeit these constraints are not explicitly stated in their design requirements.

Faller describes an echo suppressor (as opposed to an echo canceller) that reduces computational complexity by using perceptual methods and eliminating the linear adaptive filter. Disappointingly, as Faller acknowledges, and as we have also observed in our experience with echo suppressors, when the echo level is louder than the local talker’s level, the local talker signal may be suppressed. Echo suppression systems that remove the linear adaptive filter also eliminate the ability to handle real world scenarios where room acoustics, microphone and loudspeaker placement, and gain levels are not ideal—or even terrible!

The echo level may be louder than the local talker level for a variety of reasons. For example, when the local talker is soft-spoken, or is hearing-impaired or in a noisy room and requires louder than normal playback levels, the echo level may be louder than the talker level. Another scenario that may produce a louder echo is when the loudspeaker and microphone are in the same physical housing, which would be the case in a typical consumer-grade device, such as a one-piece phone.

Hands-free speakerphones do not always have the luxury of “high end” systems where loudspeakers are optimally placed and physically separated to minimize acoustical coupling with microphones. The adaptive filter serves an important purpose, especially during double-talk, by predicting and removing echo without suppressing or damaging the local talker’s voice quality. Hence, saving MIPs by removing the adaptive filter may be acceptable only in limited situations where the acoustics are (unrealistically) ideal.

Frequency domain or subband solutions, as described by Hansler and Schmidt, could be used as a starting point to overcome some of the performance and processing limitations of time domain solutions. Though those subband methods are not necessarily limited to narrowband speech signals, they are typically implemented using Fast Fourier Transform (FFT)-based analysis/synthesis methods—

---


methods whose computational complexity increases as the signal bandwidth or sample rate increases. For example, one might reasonably try to extend a narrowband AEC solution by operating at a higher sample rate and proportionally increasing the FFT size. Unfortunately, the MIPs and memory consumption increase, because the number of FFT bins increases when the FFT size increases.

The two main equations that account for most of the MIPs and memory consumption of a traditional Normalized Least Mean Square (NLMS) subband adaptive filter are the convolution equation (1) that calculates the predicted echo, \( \hat{D}_k \), and the filter adaptation equation (2) that estimates the impulse response, \( \hat{h}_{k,j} \), between the loudspeaker and the microphone.

\[
\hat{D}_k(n) = \sum_i \hat{H}_{k,j}(n)X_i(n-i)
\]

(1)

\[
\hat{H}_{k,j}(n+1) = \hat{H}_{k,j}(n) + \mu_i \frac{E_i^2X_i(n-i)}{\|X_i(n)\|^2}
\]

(2)

For each subband, \( k \), the predicted echo is calculated and the filter coefficient is updated. Clearly, as the number of subbands increases, the computational complexity increases.

**Proposed Solution**

Our solution combines the non-linear aspects of the human auditory system with the strengths of a traditional subband adaptive filter. We reduce the computational complexity by transforming the signal into a more perceptually-oriented space that is still compatible with an NLMS-based adaptive filter.

**Non-linear frequency compression**

The human auditory system maps a linear frequency scale onto a non-linear perceptual scale. Similarly, we reduce the number of FFT bins by combining them into perceptually meaningful bands such that lower frequency bins are grouped into fine resolution bands (that span small bandwidths), and higher frequency bins are grouped into coarse resolution bands (that span wider bandwidths). For example, Figure 2 shows an input-output relationship where the wider input spectrum is compressed onto a narrower perceptual scale.

![Figure 2 Compressing the input spectrum](image)

Many audio compression techniques perform this type of non-linear frequency compression that groups bins into perceptually meaningful bands in order to perform a critical-band-like analysis. This analysis helps determine auditory masking thresholds. The key difference between our method and audio compression methods is that audio compression methods perform their masking analyses in terms of power (or magnitude) and disregard the complex (real and imaginary) nature of the FFT result. (i.e., the complex FFT bins are first converted to power, then grouped into critical bands). Unfortunately, traditional subband adaptive NLMS filters that use an FFT to implement the time-to-frequency transform require complex (real and imaginary) data; the masking results from the typical power spectrum model cannot be directly used because the phase information is discarded. The challenge, then, is to apply a perceptual model but to not destroy phase information. Figure 3 shows how a traditional subband adaptive filter is modified by adding compression / decompression blocks.

In our library, time domain data is “framed” into N-sample windows with 50% overlap. That is, each successive frame of time domain data is \( N/2 \) samples apart. For each frame, the time domain samples are converted into complex frequency domain bins through an FFT. Due to the complex conjugate symmetry of the FFT result, only \( R = N/2 + 1 \) bins are...
needed. For a given frame, the $R$ bins are grouped (or effectively “compressed”) into a smaller number of $M$ perceptually relevant bands, where $M < R$. For each band, the best complex bin is selected to represent that band (for the given frame of data). Only this smaller subset of $M$ complex bins is processed through the AEC, and any subsequent Residual Echo Suppression (RES), Noise Reduction (NR) or other speech enhancement algorithms. Because the actual amount of data processed through the AEC (and other speech enhancement algorithms) is reduced from $R$ down to $M$, the overall MIPs and memory consumption decreases.

The challenge is, of course, to expand or reconstruct the smaller subset of $M$ perceptually important subbands back to the full set of $R$ bins required for the frequency-to-time conversion using the inverse FFT. Fortunately, the original, uncompressed spectrum is maintained long enough (i.e., not immediately discarded after the compression) so that it guides the reconstruction of the bins that were not processed through the AEC (and other enhancement algorithms).

For each band, the “discarded” bins (i.e., the bins not selected as the best bin for a given band and therefore not processed by the AEC) are reconstructed. This reconstruction maintains the bin’s original magnitudes and phase relationships to the best bin in the original uncompressed spectrum, but now the magnitudes and phase relationships are relative to the echo-cancelled result. The full details of this compression and decompression procedure are described in our patent application for “Sub-band processing complexity reduction” (USPTO application 20100198603).

The compression ratio, $R/M$, is the main parameter that affects the computational savings. More MIPs can be saved by reducing the number of $M$ bands that are processed by the conventional AEC. Not too surprisingly, as the compression ratio increases, the audio quality decreases. There is, nonetheless, a remarkable amount of compression that can be performed while maintaining audio quality.

Table 1 shows how the AEC computational complexity is reduced for different sample rates while using a fixed compression ratio. The “base line” complexity, when no compression is used ($R/M = 1$), is shown in row 3. The reduced complexity for a fixed compression ratio ($R/M = 2$) is shown in row 5. The computational complexity is measured in terms of the percentage of the processor used, not in MIPs. The AEC was configured to use a 200 millisecond echo tail, and the test sequences contained both single- and double-talk sections. As expected, we observed that the AEC processor load decreases by roughly the same amount as the compression ratio.

<table>
<thead>
<tr>
<th>Sample rate (kHz)</th>
<th>8</th>
<th>16</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compression ratio (R/M)</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>AEC % of processor load</td>
<td>9.18</td>
<td>17.02</td>
<td>33.93</td>
</tr>
<tr>
<td>Compression ratio (R/M)</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>AEC % of processor load</td>
<td>4.43</td>
<td>8.87</td>
<td>16.96</td>
</tr>
</tbody>
</table>

Table 1: Processor load for various sample rates and compression ratios ($R/M$).

Note that the target processor used to measure these numbers was a 462 MHz Freescale 5200 (PowerPC). Similar performance was achieved on other processors, including ARM9 / ARM11-based processors, Renesas SH-4 processors, and Intel Atom processors. Our library is portable across different processors because it is written in C (not in assembly language); it does not depend on any
specific hardware or instruction set optimizations, or on any specific operating system functions. For our target application, we increased the compression ratio until the entire library’s MIPs consumption when operating at 48 kHz was the same as when operating at 16 kHz.

Listening tests in real-world settings show that, even in challenging acoustic environments, we have maintained audio quality, including full-duplex performance.

Conclusions

We have developed novel methods for reducing the computational complexity and memory requirements for Acoustic Echo Cancellation. Our solution has been validated and will be deployed (summer 2010) in existing AEC systems that allow upgrading from wideband processing to full-band processing, maintaining audio quality without additional MIPs or memory costs.

The techniques we have employed also make it possible to deliver true telepresence—a combination of wider bandwidth speech (super-wideband) and three-dimensional spatial sound (stereo telephony)—on processors used in everyday terminals, such as laptops and conference speakerphones. These telepresence systems dramatically increase the degree of realism in speech communications, and could easily become the next “killer app” for telephony.

The catch has been, however, that the computational demands created by multi-channel AEC, which is required for three-dimensional spatial sound, can be prohibitive. Using traditional AEC techniques the processor demands will increase by a factor of $C \times D$, where $C$ is the number of reference signal channels and $D$ is the number of microphone channels. This means that the combination of stereo AEC (two references times two microphones creating four times the load) and wider bandwidth speech (twice the load) will require eight times ($4 \times 2 = 8$ times the load) more processor resources than traditional telephony—placing telepresence out of reach for most terminals. The techniques described in this paper make it possible to now deliver telepresence communications, not just high-end telepresence systems costing upwards of US$30,000, but on everyday terminals. It is exciting to think about a future with telepresence for everyone!

Acknowledgements

The authors wish to thank their QNX (Wavemakers) colleagues for supporting the research and development of this project. Additionally, some of the ideas and methods described in this paper have been submitted in a patent application and are patent pending.

References


About QNX Software Systems

QNX Software Systems is the leading global provider of innovative embedded technologies, including middleware, development tools, and operating systems. The component-based architectures of the QNX® Neutrino® RTOS, QNX Momentics® Tool Suite, and QNX Aviage® middleware family together provide the industry’s most reliable and scalable framework for building high-performance embedded systems. Global leaders such as Cisco, Daimler, General Electric, Lockheed Martin, and Siemens depend on QNX technology for vehicle telematics and infotainment systems, industrial robotics, network routers, medical instruments, security and defense systems, and other mission- or life-critical applications. The company is headquartered in Ottawa, Canada, and distributes products in over 100 countries worldwide.

www.qnx.com

© 2010 QNX Software Systems GmbH & Co. KG, a subsidiary of Research In Motion Ltd. All rights reserved. QNX, Momentics, Neutrino, Aviage, Photon and Photon microGUI are trademarks of QNX Software Systems GmbH & Co. KG, which are registered trademarks and/or used in certain jurisdictions, and are used under license by QNX Software Systems Co. All other trademarks belong to their respective owners. 302167 MC 411.83