14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

A HIGH PERFORMANCE, LOW LATENCY, LOW POWER AUDIO PROCESSING SYSTEM FOR WIDEBAND SPEECH OVER WIRELESS LINKS

Etienne Cornu[1], Alain Dufaux[2], and David Hermann[1]

[1]AMI Semiconductor Canada, 611 Kumpf Drive, Unit 200, Waterloo, Ontario, Canada N2V 1K8
[2]AMI Semiconductor Switzerland, Champs-Montants 12a, 2074 Marin, Switzerland
phone: +1 (519) 884-9696, fax: +1 (519) 884-0228, email: etienne_cornu@amis.com
web: www.amis.com

ABSTRACT

In this paper we present an audio processing system optimized for bi-directional wideband speech processing over wireless links. The system's architecture features a DSP core, a WOLA filterbank coprocessor and an I/O processor. We show how speech encoding, decoding and enhancement algorithms can be deployed simultaneously on this platform. The resulting system may include a combination of features that are usually not found when standard codecs are deployed on standard DSPs in wireless headsets. The algorithms can be combined through efficient re-use of the filterbank coprocessor for oversampled and critically sampled subband processing. Such a system may include features such as low latency, ultra-low power consumption, resilience to background noises and music and the possibility to add forward error correction for graceful degradation in difficult RF environments.

1. INTRODUCTION

As emerging gaming and VoIP applications support wideband speech (up to 7 kHz audio bandwidth), speech processing techniques used for normal telecommunication are no longer sufficient to provide the required signal quality. End users expect noticeably increased quality from wideband speech communications. The use of wideband speech in wireless speech transmission also requires the use of a specialized audio codec since many wireless speech communication links are only designed for narrowband audio transmission.

For example, in Bluetooth, speech is typically sampled at 8 kHz using 16 bits and compressed to 64 kbits/sec. In order to maintain low power consumption and take advantage of the existing transport protocols at 64 kbits/sec, more complex data compression and transmission techniques must be used for wideband speech. Wideband speech is typically sampled at 16-bits and 16 kHz, therefore a minimum 4-to-1 compression ratio for the speech signal is required. A number of codecs are able to provide this compression ratio for wideband speech. However, compression ratio is not the only factor in choosing a codec for wireless transmission of wideband speech. Other issues to consider are robustness to packet loss, low latency in encoding and decoding and low complexity for low power consumption. The codec must also be robust in noise of all types, and should minimize distortion when coding music since these are both common in gaming applications that feature musical soundtracks and loud audio effects. It is also important to consider how the codec will integrate with speech quality enhancement algorithms such as noise reduction and dynamic range compression. Thus, a complete wideband speech processing system will require a robust codec which can be easily integrated with additional algorithms in order to provide improved sound quality in noisy environments. The codec should also be robust in the presence of lossy wireless transmissions, by using features such as variable bit-rates or side channels with smaller bit allocations to account for isolated packet drops.

Although existing wideband speech codecs (ITU-T G.722.1 [1], G.722.2 [2], for example) usually provide good performance, they may not be suitable for low latency, low power applications because of relatively long delays, poor speech coding performance in certain noisy conditions, excess distortion when coding music or high computational resource requirements.

In this paper we describe an audio processing system whose architecture is optimized for low-latency, low-power audio processing that can encode, decode and improve the quality of wideband speech in wireless applications. Codecs and other speech processing algorithms can be deployed in the subband domain, which has many advantages. For example, subband-based codecs can provide general coding ability like other transform-domain audio codecs, but can still be optimized for speech signals since speech is still the primary signal of interest. The subband approach also provides a useful framework for frequency-domain noise reduction and multi-band dynamic range compression [3]. In this application, the use of a weighted overlap-add (WOLA) filterbank coprocessor in an audio processing DSP provides a means to implement this subband processing while introducing very low latency and achieving ultra-low power consumption [4].

The content of this paper is organized as follows. Section 2 describes the architecture of the audio processing system and its integration in wireless devices. In section 3 we present the techniques required to be able to use the filterbank for subband encoding/decoding as well as other subband speech enhancement processing. In section 4 we describe how bi-directional encoding, decoding and signal processing are deployed on the audio processing system. In section 5 we present some characteristics of the algorithms deployed on the audio processing system, and finally in section 6 we present a summary and describe possible future work.

2. AUDIO PROCESSING SYSTEM

The audio processing system used in this application is designed for ultra-low power applications such as hearing aids and wireless headsets [4]. The digital part of the audio processing system consists of three major components: a weighted overlap-add (WOLA) filterbank coprocessor, a 16-bit fixed-point DSP core, and an input-output processor (IOP). These three components run in parallel and communicate through shared memory and interrupts. The parallel operation of these components allows for the implementation of complex signal processing algorithms with low system clock rates and low resource usage. The system is particularly efficient for subband processing since the configurable WOLA filterbank coprocessor efficiently computes the analysis and synthesis filterbank operations while leaving the DSP core free to perform the other algorithm calculations. The WOLA filterbank coprocessor also features a vector multiply function that can be used for applying gains to the incoming and outgoing signals in the frequency domain without intervention of the DSP core.

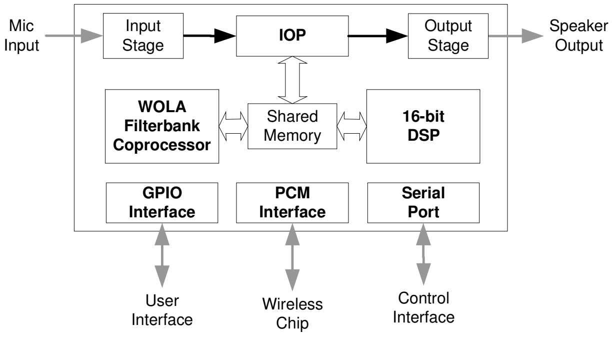

Figure 1 – Audio processing system architecture

The flow of the incoming and outgoing audio samples is managed by the IOP. It coordinates the data flow between the FIFO memory and the A/D and D/A converters without intervention from the DSP core. The DSP core does not begin processing of the samples until a specific number of samples have been stored in the FIFO. The IOP raises a periodic interrupt (at a rate called the tick rate) to signal the DSP core that a new block of R samples has been written to the FIFO, and that a new block of samples must be ready in the FIFO to be read for output. In many wide-band speech applications, R is typically 16, resulting in a tick rate of 1 millisecond at a sampling frequency of 16 kHz.

In addition to the three processing units that operate in parallel, the audio processing system also features a PCM interface for communication with a wireless chip, a serial port for control interfacing and a number of general-purpose input/output (GPIO) pins for other interfacing as shown in Figure 1.

The FIFO memory and the filterbank coprocessor are designed so that they can process two signal streams in parallel. These streams may represent a unidirectional stereo signal or, as in this application, a bi-directional mono stream. In applications running on the audio processing system, most of the actual processing of these two signal streams occurs in the subband-domain data produced by the filterbank coprocessor.

3. SUBBAND PROCESSING

Subband processing is a useful methodology for both speech enhancement and audio coding. Classical speech enhancement algorithms can be mapped to an oversampled subband processing scheme in order to produce high-fidelity outputs. This oversampling allows greater freedom in the analysis and synthesis filter design, as well as allowing these filters to be shorter in length which ultimately reduces the group delay through the filterbank. Typically, complex-modulated (i.e., DFT-based) filterbanks are used.

Subband coding is also a popular method for low-resource audio coding such as that used in the Bluetooth Advanced Audio Distribution Profile (A2DP) [5]. In order to minimize the amount of data to be quantized, the filterbanks used in these audio coding applications generally provide real-valued, critically-sampled signals in the subbands. In other words, the channel filters are real and the subbands are maximally decimated. Typically, cosine-modulated filterbanks are used.

The WOLA filterbank coprocessor readily provides the oversampled subband signals necessary for high-quality speech enhancement algorithms, as it realizes an oversampled, N-channel complex-modulated filterbank, (hereafter called a DFT filterbank), which provides complex data in its subbands. For audio coding applications requiring a cosine modulated filterbank, real-valued, critically sampled signals can be obtained by transforming the N channels of the WOLA filterbank into the M=N/2 channels of a corresponding cosine-modulated filterbank. This ability to re-use the WOLA filterbank architecture for both complex, oversampled processing and real-valued, critically sampled processing allows the integration of subband coding and subband speech enhancement algorithms.

A similar complex to real-valued filterbank conversion was originally explained in [6] in the context of the Bluetooth Subband codec. It is further developed here into a form independent of the number of channels (M) and the length of the prototype filter (L). This is important for the design of low-delay codecs, where the length of the prototype filter can be optimized by taking advantage of different filterbank configurations and improved prototype window design.

The transformation from the subbands of the DFT filterbank to the subbands of the cosine-modulated filterbank can be derived by writing and comparing the z-domain expressions for the k-th subband in both filterbanks. In the following, $x(n)$ is the input signal to the filterbank, $h_k(n)$ is the filter for channel k, R is the decimation factor in the filterbank (which always equals M for a critically sampled cosine-modulated filterbank), and $X_k(m)$ is the R-times decimated subband centred at frequency $\omega_k = (k+0.5)\pi/M$. Note that m is the decimated time index in the subbands. Figure 2 shows the DFT filterbank analysis for the k-th channel (k=0,…, N-1) [11].

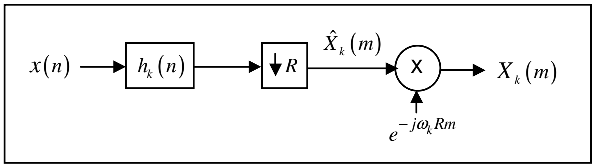

Figure 2 – DFT filterbank analysis

In the case of the DFT filterbank, the filter impulse response $h_k(n)$ takes the form (1), with $h(n)$ the low-pass prototype filter. Then the z-transform (2) of $\hat{X}_k(m)$ can be evaluated, involving the transfer function of the downsampling block. $X(z)$ and $H(z)$ are the respective z-transforms of $x(n)$, the input signal, and $h(n)$, the prototype filter impulse response. For now, let us discard the subsequent complex modulation leading to the final output subband $X_k(m)$.

$$h_k(n) = h(n)\ \exp\left[j\frac{2\pi}{N}\left(k+\frac{1}{2}\right)\left(n-\frac{L}{2}\right)\right] \tag{1}$$

$$\hat{X}_k(z) = \frac{1}{R}\sum_{l=0}^{R-1} e^{-j\omega_k \frac{L}{2}} X\left(z^{\frac{1}{R}} e^{-j\frac{2\pi l}{R}}\right) H\left(z^{\frac{1}{R}} e^{-j\frac{2\pi l}{R}} e^{-j\omega_k}\right) \tag{2}$$

In the case of the cosine-modulated filterbank, the analysis operations are the same as in Figure 2, but the final complex modulation is absent. Furthermore, the filter impulse response $h_k(n)$ takes a different form (3). The z-transform of the output subband $X_k(m)$ is expressed in (4), for a particular channel k (k=0,…, M-1).

$$h_k(n) = 2\ h(n)\ \cos\left[\frac{\pi}{M}\left(k+\frac{1}{2}\right)\left(n-\frac{L}{2}\right)+\frac{\pi}{4}(-1)^k\right] \tag{3}$$

$$\begin{aligned} X_k(z) &= \frac{1}{R}\sum_{l=0}^{R-1} e^{j\frac{\pi}{4}(-1)^k} e^{-j\omega_k\frac{L}{2}} X\left(z^{\frac{1}{R}} e^{-j\frac{2\pi l}{R}}\right) H\left(z^{\frac{1}{R}} e^{-j\frac{2\pi l}{R}} e^{-j\omega_k}\right) + \ldots \\ &+ \frac{1}{R}\sum_{l=0}^{R-1} e^{-j\frac{\pi}{4}(-1)^k} e^{j\omega_k\frac{L}{2}} X\left(z^{\frac{1}{R}} e^{-j\frac{2\pi l}{R}}\right) H\left(z^{\frac{1}{R}} e^{-j\frac{2\pi l}{R}} e^{j\omega_k}\right) \end{aligned} \tag{4}$$

Assuming the same prototype $h(n)$, the subband expressions (2) and (4) can now be compared. A careful observation shows that the M subbands of the cosine-modulated filterbank can be obtained by performing the following transformation to the channels of the DFT filterbank. In this expression, the range of the channel index is k=0,…, M-1:

$$\begin{aligned} X_k^{COS}(m) &= e^{j\frac{\pi}{4}(-1)^k} \hat{X}_k^{DFT}(m) + e^{j\frac{\pi}{4}(-1)^{2M-1-k}} \hat{X}_{2M-1-k}^{DFT}(m) \\ &= 2\ Real\left\{e^{j\frac{\pi}{4}(-1)^k} \hat{X}_k^{DFT}(m)\right\} \end{aligned} \tag{5}$$

Now, taking into account the complex-modulation in Figure 2, this shows that $\hat{X}_k(m) = e^{j\omega_k Rm} X_k(m)$, which leads to the final transformation (6), for k=0,…, M-1:

$$X_k^{COS}(m) = 2\ Real\left[X_k^{DFT}(m)\ e^{j\left[\omega_k Rm + \frac{\pi}{4}(-1)^k\right]}\right] \tag{6}$$

This expression can be applied to the output of the analysis operation performed by the WOLA filterbank coprocessor, to convert it into a cosine-modulated filterbank.

At synthesis, an opposite transformation can be similarly derived by comparing the z-transforms of the time-domain re-synthesized signals for both filterbanks. Such a development shows that if the cosine-filterbank channels are modified according to equation (7), a subsequent DFT filterbank synthesis will rebuild the same output signal as a cosine filterbank synthesis. For k=0,…, M-1 :

$$X_k^{I\_DFT}(m) = X_k^{COS}(m)\ e^{-j\left[\omega_k Rm + \frac{\pi}{4}(-1)^k\right]} \tag{7}$$

Hence, transformation (7) must be applied before the synthesis operation performed by the WOLA filterbank coprocessor. $X_k^{I\_DFT}(m)$ becomes the input to the WOLA filterbank synthesis.

Within the context of our application, it is important to note that these expressions only involve phase shift operations. The amount of phase shift in every band varies cyclically with the block index m. Hence, for calculation efficiency, the shift values can be stored in a table and applied as complex gains, an operation that can be performed efficiently by the WOLA filterbank coprocessor.

The coprocessor also takes advantage of the fact that the computation of the filterbank for two channels (one for noise reduction and one for coding, for example) can be simplified by well-known methods for computing the DFT transform of two real-valued sequences using a single complex DFT [7]. The samples for each channel are stored in the FIFOs of the DSP system in interleaved format and the same prototype filter is applied to each sequence. The resulting interleaved sequence can be viewed as a set of complex-valued pairs, where the real part corresponds the first channel and the imaginary portion corresponds to the second channel. A single complex-valued DFT is then used to produce the filterbank analysis results for two channels, after a suitable separation step as described in [7]. This saves memory and computation when compared to performing two independent filterbank analysis operations. An analogous procedure can be used to efficiently synthesize two channels that use the same filterbank configuration.

4. CODEC AND ALGORITHM INTEGRATION

The audio processing system's architecture and its ability to process two channels in parallel allows subband-based codecs and algorithms to be integrated in an efficient way. When the audio processing system is deployed in a wireless headset, these algorithms may typically include noise reduction on the transmit channel and dynamic range compression on the receive channel. Processing of these two channels requires the execution of a number of operations on the WOLA filterbank coprocessor. For the transmit channel, these operations include: a first transformation of the time-domain microphone signal into the frequency domain for processing by the noise reduction algorithm, the application of gains calculated by the algorithm, transformation back into the time-domain, a second transformation into the frequency domain for the encoder and finally the application of gains to perform the phase shift operation described in the previous section. The signal is then ready to be processed by the encoder. On the receive channel, the operations performed by the WOLA filterbank coprocessor include: a first gain application to perform the inverse phase shift, a second gain application for volume control, and finally a transformation of the signal back into the time domain where it can be played on the loudspeaker. Figure 3 illustrates how these WOLA filterbank coprocessor operations are performed for every set of *R=16* samples and executed in parallel with processing performed on the DSP core.

As illustrated in Figure 3, the microphone signal stored in the input FIFO by the IOP first goes through a set of transformations that results in noise-reduced time-domain signal in the output FIFO. This signal is copied back to the input FIFO and transformed again into the frequency domain, this time for processing by the wideband speech encoder in a critically sampled manner. After the filterbank conversion (phase shift) and the actual encoding and compression are performed, frames of are then sent to the RF chip via the PCM interface.

The incoming signal, in the form of frames of encoded speech, reaches the audio processing system via the PCM interface. The wideband speech decoder extracts the subband data from these frames and stores them so that the WOLA filterbank coprocessor can transform the filterbank into a complex modulated filterbank, apply a gain (set by the user using volume control buttons on the headset), transform the signal into the time domain and place it in the output FIFO, where it will be eventually played back on the speaker.

As a result of this architecture, the outgoing signal is processed in two ticks, one for noise reduction and one for encoding. A single tick is needed for processing the incoming signal.

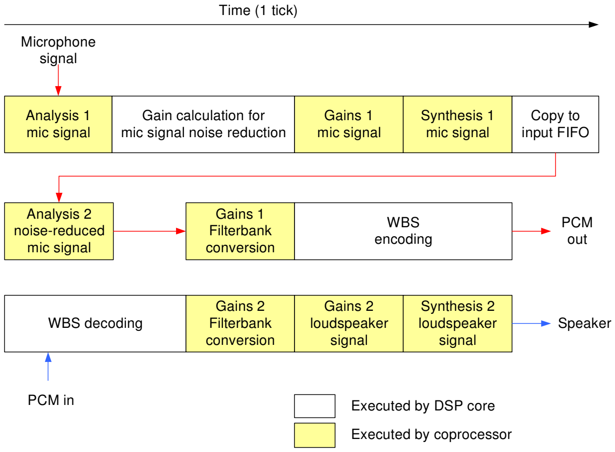
Time (1 tick)
Microphone signal
Analysis 1 mic signal
Gain calculation for mic signal noise reduction
Gains 1 mic signal
Synthesis 1 mic signal
Copy to input FIFO
Analysis 2 noise-reduced mic signal
Gains 1 Filterbank conversion
WBS encoding
PCM out
WBS decoding
Gains 2 Filterbank conversion
Gains 2 loudspeaker signal
Synthesis 2 loudspeaker signal
Speaker
PCM in
Executed by DSP core
Executed by coprocessor

Figure 3 – Signal flow and algorithm scheduling

5. SYSTEM EVALUATION

The audio processing system and the integration framework shown in Figure 3 lend themselves to a range of subband-based codecs and speech processing algorithms. The following sections describe sample modules that can be integrated into the framework and deployed in Bluetooth wireless headsets.

a. Codec. Codec functionality required for wireless transmission of wideband speech can be implemented by coding each subband using a simple adaptive PCM quantization scheme. One such scheme is a modified version of the one-word memory adaptive quantization scheme of [8]. The adaptation occurs in each subband using quantizer step sizes and adaptation multipliers that are designed for processing speech signals. This quantization scheme provides reasonable subband coding performance for speech signals with a minimum of computational resources. For example, a typical implementation of a codec may produce a nominal bitstream of 48 kbps and encode the frequency range from 0 – 7500 Hz.

Informal tests were done to evaluate the quality of this wide-band speech codec. Using standard objective speech quality measurements, it achieves an average predicted MOS score of 3.6. Additional tests with a wide variety of background noises showed that the codec could also handle non-speech environmental sounds without serious degradation.

When the nominal codec bitrate is only 48 kbps, there is additional bandwidth in a typical 64 kbps link to perform forward error correction that can mitigate the effect of lost or damaged audio packets in poor RF environments. Good results can be obtained by including selectively encoded sections of the spectrum for this purpose.

b. Noise reduction. An example of a noise reduction algorithm that can be deployed on the system is a low-resource subband noise reduction algorithm based on the MMSE spectral noise reduction rule of Ephraim & Malah [9]. This algorithm has been tested in real-life acoustic situations and shown to produce up to 16 dB of SNR improvement in both white and "convention" type noise [10].

c. Dynamic Range Compression. Gain adjustment algorithms including multi-band dynamic range compression or even simple volume control can be implemented using the WOLA filterbank coprocessor by applying gains to the subband signals before synthesis [3].

The encoding and decoding algorithms described above require approximately 4.4 MIPS in total on the DSP core. This is about an order of magnitude less than G.722.2. As previously described, all the filterbank processing occurs on the WOLA filterbank coprocessor. This filterbank coprocessor provides another 5 MIPS worth of equivalent processing power. The noise reduction algorithm from above requires approximately 5 MIPS on the DSP core. Gain application is performed on the WOLA filterbank coprocessor, so the combined set of algorithms can be implemented on the DSP system with a system clock less than 10 MHz. The total power consumption for this configuration is less than 6 mW at a 1.8 Volt power supply.

In a typical Bluetooth setting, 30-byte data packets are transmitted every 4 milliseconds. Using the framework described in Section 4, block processing for noise reduction and encoding takes 2 milliseconds. An additional 6 milliseconds must also be taken into consideration for the filtering that takes place as part of the filterbank analysis. Block delay for decoding is 1 millisecond and the delay for the filtering associated with synthesis is 6 milliseconds, resulting in a total latency through the audio processing system of 19 milliseconds, which is about half of the low complexity codec G.722.1. If a forward error correction scheme is deployed, additional delays would normally be incurred. This additional delay is nominally the time length of one additional packet (4 ms).

6. CONCLUSIONS AND FUTURE WORK

We have presented an audio processing system and an integration framework for the processing of wideband speech in wireless headsets. Deployment of an encoder, decoder and other subband based algorithms results in a system with low latency and low power consumption that provides good signal quality compared with existing solutions. The system proposed by this paper includes noise reduction offering up to 16 dB of SNR improvement and a simple subband audio codec with an average predicted MOS score of 3.6. These algorithms can be deployed with a 10 MHz system clock and requires less than 6 mW of power consumption with a 1.8 V supply. Future work will include integrating other algorithms such as echo cancellation in the framework in order to allow a wide variety of mechanical designs. This will also include decreasing the total latency by improving the filter design of the filterbank.

REFERENCES

[1] ITU-T Recommendation G.722.1, "Coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss".

[2] ITU-T Recommendation G.722.2 (2002) "Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB)".

[3] T. Schneider, R. Brennan, "A Multichannel Compression Strategy For A Digital Hearing Aid", *Proc. ICASSP 1997*.

[4] R. Brennan and T. Schneider, "Filterbank Structure and Method for Filtering and Separating an Information Signal into Different Bands, Particularly for Audio Signal in Hearing Aids", United States Patent 6,236,731. WO 98/47313. April 16, 1997.

[5] Advanced Audio Distribution Profile. Bluetooth Audio Video Working Group. 2002.

[6] D. Hermann et. al., "Low-Power Implementation of the Bluetooth Subband Audio codec", *Proc. ICASSP 2004*. Montreal, Canada, May 2004

[7] H. Sorenson et. al., "Real-Valued Fast Fourier Transform Algorithms", *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vo. 35, No. 6, June 1987.

[8] N. S. Jayant and P. Noll, *Digital Coding of Waveforms*, Prentice Hall, 1984.

[9] Y. Ephraim, D. Malah, "Speech Enhancement Using a Minimum Mean-Square Log-Spectral Amplitude Estimator", *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-33 (2), pp. 443-445. 1985.

[10] K. Tam and E. Cornu, "System architecture for audio signal processing in headsets", *Global Signal Processing Expo (GSPx) Proceedings 2005*, Santa Clara, USA, Oct. 2005.

[11] N. J. Fliege, *Multirate Digital Signal Processing*, Wiley, 1994