To appear in *ICSPAT 2000 Proceedings*, October 16-19, 2000, Dallas, TX.

An Ultra Low-power Miniature Speech CODEC at 8 kb/s and 16 kb/s

Robert Brennan, David Coode, Dustin Griesdorf, Todd Schneider

Dspfactory Ltd.
611 Kumpf Drive, Unit 200
Waterloo, Ontario, N2V 1K8
Canada

Abstract

This paper describes a CODEC implementation on an ultra low-power miniature application specific signal processor (ASSP) designed for mobile audio signal processing applications. The CODEC records speech to and plays back from a serial flash memory at data rates of 16 and 8 kb/s, with a bandwidth of 4 kHz. This CODEC consumes only 1 mW in a package small enough for use in a range of demanding portable applications. Results, improvements and applications are also discussed.

1. Introduction

Speech coding is ubiquitous. Increasing demands for portability and fidelity coupled with the desire for reduced storage and bandwidth utilization have increased the demand for and deployment of speech CODECs.

However, many current devices using speech CODEC technologies consume enough power to put inconvenient limits on battery lifetime. As well, CODECs are sometimes too large for mobile applications.

The CODEC presented here is an application of the programmable SmartCODEC platform. In this implementation, the memory chip consumes most (98%) of the power, and the entire package is extremely small.

2. System Overview

The CODEC was designed by interfacing the ultra low-power SmartCODEC hardware platform with a 32 Mbit serial flash. The SmartCODEC platform consists of an efficient, block-floating point, oversampled Weighted OverLap-Add (WOLA) filterbank, a software-programmable dual-Harvard 16-bit DSP core, two high fidelity 14-bit A/D converters, a 14-bit D/A converter and a flexible set of peripherals [1]. The system hardware architecture (Figure 1) was designed to enable memory upgrades. Removable memory cards or more power-efficient memory could be substituted for the serial flash memory.

The CODEC communicates with the flash memory over an integrated SPI port that can transfer data at rates up to 80 kb/s. For this application, the port is configured to block transfer frame packets every 14 ms.

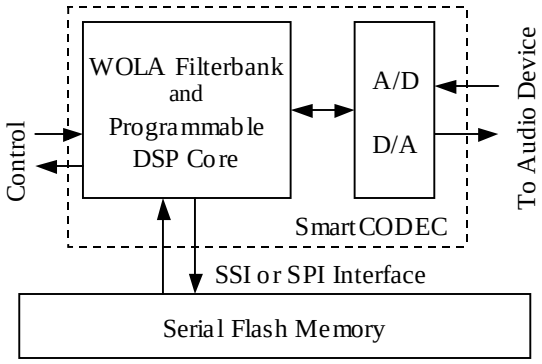

Figure 1. System Block Diagram

The SmartCODEC also has 2 UARTs. The CODEC uses one UART for control signals (play / stop / record / time). The other is used for an integrated, on-chip debugging port.

In its most compact incarnation (Figure 2), the SmartCODEC platform measures 6.5 x 3.5 x 2.5 mm [2]. The flash memory, minimal interfacing hardware and a battery (when no

other power source is present) occupy additional space.

Figure 2. Relative size of SmartCODEC platform

A non-volatile 32-Mbit serial flash memory is currently employed to store the speech. This chip stores up to 56 minutes of speech at 8 kb/s and measures 20.2 x 7.5 x 1.2 mm [3].

The power consumption of the complete system is 55 mW for recording and 41 mW for playback. The SmartCODEC platform itself consumes less than 1 mA at 1 volt, only 2% of the total system power. The remainder of the power is consumed by the flash memory.

3. Coding Algorithm

The CODEC design follows traditional Sub-Band Coding (SBC) methods [4]. Speech is sampled at 16 kHz and analyzed into 16 complex frequency bands via a 32-point FFT and a 256-point prototype low pass filter. The filterbank is 2 times oversampled. This results in 32 words (8 kHz) of complex data, 2 words for each 500 Hz band.

The incoming bit rate of 16 kHz monaural speech is 256 kb/s. After analysis the oversampling increases this rate to 512 kb/s. The rate is subsequently reduced to critical sampling by careful decimation of the complex frequency data. The analysis data at this point is in the form of a DCT [5]. Discarding the frequencies above 4 kHz further reduces the data rate to 128 kb/s.

The remaining 8 words of data represent 1 ms of speech data. These 1 ms blocks are buffered into groups of 14. Each group comprises a 14 ms frame of speech data. Each frame includes its own bit allocation over all 8 bands and its own noise floor.

Figure 3 shows the design of the software data structures and processing modules for encoding speech. Each storage element is double-buffered, allowing a set of data to remain static during a processing frame. The other half of a buffer actively accumulates data in real-time.

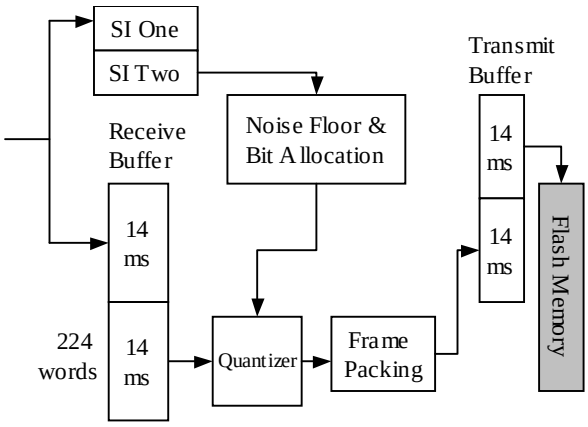

Figure 3. Speech Encoder

As the Receive Buffer accumulates 14 blocks of data, a record of the maximum level in each band is kept as the Spectral Image (SI). The Spectral Image reveals the relative magnitude of all 8 frequency components. A parallel algorithm calculates a flat noise floor and a near-optimal bit allocation across 8 bins based on this Spectral Image. This data configures the quantizer to encode the spectral levels from the Receive Buffer so that distortion is minimized. The bits represent the level as a multiplier of the noise floor.

Frames of compressed data are packed with the noise floor first followed by the bit allocation and finally the quantizer output over 14 blocks. These frames are then sent to the flash memory via the SmartCODEC's onboard SPI port for storage in the serial flash memory.

The noise floor is packed by storing a 4-bit exponent and a 4-bit mantissa, with two virtual bits (Figure 4). The "virtual bits" are two known bits which are not transmitted because they are constant. The sign of the noise floor (one virtual bit) is always positive and the most significant bit in the mantissa (other virtual bit) will always be 1. This noise floor

compression was found to be extremely accurate, since the noise floor has to rise up over 30 dB before even 1 LSB quantization error occurs. Accuracy is important since the quantized noise floor is the basis for decoding each level.

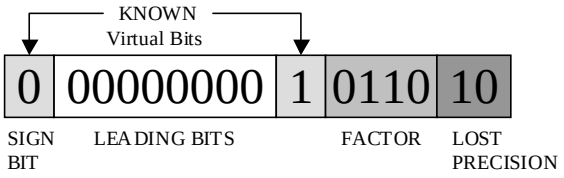

Figure 4. Noise Floor Encoding (Virtual Bits)

In contrast to the encoding process, the decoding operation (Figure 5) is simple. Every 14 ms in playback mode, a compressed frame is fetched from the serial flash. The decoder first unpacks the noise floor, then the bit allocation over all 8 bands. From the bit allocation information, the decoder can properly parse the remaining bits in the frame (the encoded speech data levels from the quantizer). Multiplying the encoded levels by the noise floor recalculates the original levels in each band, which are sent to the SmartCODEC synthesis buffer every millisecond.

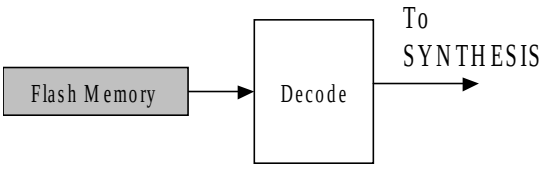

Figure 5. Decoder

The system polls for control commands (play / record / stop) after each frame is processed. There is an internal state machine, whose states consist of sleeping, skipping to a certain time index in the flash, playing (decoding), recording (encoding) or doing both encode and decode simultaneously. When encoding and decoding in parallel, the speech is still recorded to flash, but the decoded output is also present at the audio output.

Implementation has proven that the SmartCODEC platform has more than enough processing power to run encoding, decoding and write to the serial flash simultaneously.

4. Results

At 16 kb/s the CODEC was informally judged to be 'good' by both experts and those unfamiliar to speech coding. At 8 kb/s it was judged to be 'suitable for low-fidelity communications'.

Due to the lossy nature of the compression algorithm, some distortions are expected. Coded speech has an associated "gurgling" quality, especially at vowel onset. Informal listening tests have compared this CODEC's performance at 16 kb/s to other coders with a speech quality MOS of 3.4.

Segmental SNR measurements were done on the CODEC's floating-point simulation at rates of 8, 16 and 32 kb/s (Table 1). Segmental SNR gives a good relative measure of coder performance where all distortions come from the same fundamental algorithm. The 32 kb/s simulation produced "transparent" results. The results of the fixed-point implementation on the SmartCODEC platform matched the results of the floating-point simulation.

Table 1. Segmental SNR measurements for the CODEC at various bit rates

Bit Rate	*Seg. SNR*
8 kb/s	15.28
16 kb/s	22.81
32 kb/s	33.24

Since the algorithm used is the same in all cases (other than the available bits), these measurements of MSE show how the CODEC distorts less at higher data rates.

A coherence measurement between the original signal and the CODEC output (Figure 6) shows the relative distortions in each band. Coherence shows the energy in the output signal that is linearly related to the input. The "Ratio" (Figure 7) was calculated by Ratio=$(1-1/R)^2$ in each band, where R=(signal energy)/(quantization noise floor). The un-

usual shape of this speech signal's overall spectrum (with an uncommon dip at 2 kHz) reveals the high correlation between energy in a band and distortion introduced in that band. The distortion for this speech signal at 2 kHz accounts for a majority of audible distortion. Furthermore, human hearing is extremely sensitive at 2kHz, making any distortions all the more audible in this specific test case.

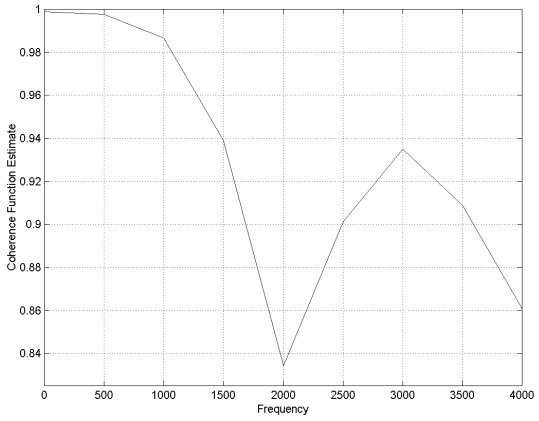

Figure 6. Coherence by 500Hz Band at 16 kb/s

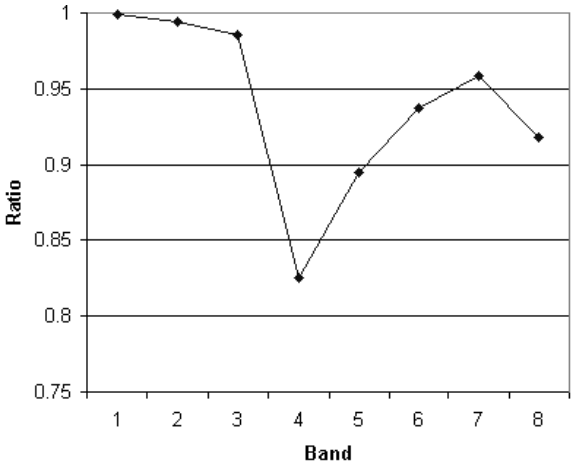

Figure 7. Ratio of Speech Signal Energy to Noise Floor in Each Band at 16 kb/s

Analysis shows that only 35% of available processing cycles are used while simultaneously encoding and decoding at the default core clock frequency of 1.28MHz. The SmartCODEC is exceptionally efficient because the analysis and synthesis steps are performed (on the WOLA filterbank) in parallel with the DSP core, as it runs the coding algorithm.

The psychoacoustic model added to this CODEC produced results that did not meet with expectations. The model used the output of the WOLA filterbank directly, instead of using a higher resolution FFT. Our results showed that 500Hz bands did not give sufficient frequency resolution to generate meaningful masking thresholds. Thus, there is no psychoacoustic model included in the current version.

5. Improvements

Several improvements have been made to the basic coding algorithm.

A smoothing technique was applied to the noise floor to limit how much it changed between frames. This feature is useful for preventing noise floor fluttering. The SNR is decreased with smoothing, but the perceived quality was judged to be better by trained listeners.

Experimental attempts were made to statically equalize the noise floor to match the energy spectrum of the speaker. The results of this shaping produced negligible improvements in overall distortion, and required a priori knowledge of the input spectrum.

Research suggests that 1-4 kHz is the "sweet spot" for human hearing, and that quantization noise should be shifted out of this frequency range [6]. A static shaping of the noise floor to weight bands 3-8 has the potential to improve perceived quality.

Compressing frames whose energy falls under a noise threshold with silence is possible. A reserved escape word could be transmitted at the beginning of a frame to indicate silence, and the next frame would start at the subsequent word. This introduces a variable bit rate which is guaranteed to be less than or equal to the fixed rate.

Simulations of the CODEC at 32 kb/s showed that there were enough bits available to ex-

pand the bandwidth to 8kHz and obtain higher fidelity recordings.

Another CODEC is currently under development that will incorporate a higher bit rate and a psychoacoustic model very similar to MPEG [7]. This CODEC will share the same features of low power and miniature size of the SmartCODEC platform, but will be able to encode music and speech with high fidelity.

6. Applications

The CODEC can be used in a number of ultra low-power portable applications. For example, it may be used as a speech-recording accessory for a personal digital assistant (PDA). Given its small size and ultra-low power consumption, it is also suitable for use in 3G cellular telephones or other wireless applications such as two-way pagers that incorporate voice attachments. It can also be useful in short distance, voice-band wireless communications.

The CODEC presented here illustrates the potential of the SmartCODEC for use in portable devices including wireless applications. The low-bit rate is ideal for wireless transmission as power is saved both in the encoding and transmission. High fidelity reproduction of recorded speech is possible using the platform. The small size and ultra low-power consumption make this an ideal solution for adding voice recording/playback capabilities to existing devices. New devices can incorporate the SmartCODEC hardware and software directly for further power savings. The inclusion of external memory ensures ample and upgradeable storage for a variety of applications.

In wireless devices, voice attachments and instant voice messages can be saved and played back through the memory chip. The advent of smaller, more efficient memories will complement the current system and expand its capabilities.

7. Conclusions

This paper presents an ultra low-power miniature CODEC implemented on the SmartCODEC platform. The CODEC portion of the system (not including flash memory) measures only 6.5 x 3.5 x 2.5 mm and consumes less than l mW. This CODEC is suitable for a number of personal audio and wireless applications.

The CODEC is pleasant to use at 16 kbps (an estimated speech quality MOS of 3.4 from informal tests) and remains useful at 8 kbps, yielding nearly one hour of recorded speech on a single 32 Mbit flash memory. This application uses less than half the computational capacity of the SmartCODEC platform.

References

[1] Schneider, T., Brennan R.L., “An Ultra Low-Power Programmable DSP System for Hearing Aids and Other Audio Applications”, *Proc. ICSPAT 1999*, Orlando, FL, November 1999.

[2] Brennan, R.L., Schneider, T., “A Flexible Filterbank Structure for Extreme Signal Manipulations in Digital Hearing Aids,” *Proc. ISCAS-98*, Monterey, CA, June 1998.

[3] ATMEL Corporation, *32-Megabit 2.7V Serial DataFlash AT45D321*, Data Sheet AT45D321.

[4] Jayant, N.S. & Noll, P., *Digital Coding of Waveforms*, Prentice-Hall Inc., 1984.

[5] Fliege, N. J., “Modified DFT Polyphase SBC Filter Banks with Almost Perfect Reconstruction”, *Proc. ICASSP-94*, Adelaide, Australia, pp. III149-III152.

[6] Dipert, Brian, “Digital Audio Breaks the Sound Barrier”, *EDN Magazine*. July 20, 2000, pp.71-90.

[7] ISO/IEC JTC1/SG29/WG11 MPEG, “Information Technology - Generic coding of moving pictures and associated audio information – Part 3: Audio”, IS13818-3 1994 (“MPEG-2”).