

The Serious Business of Sound for Toys (or how I went from HLA to Amazing Amy)

by Frank Ostrander

Presented at the Los Angeles AES Chapter Meeting, April 25, 2000

Introduction

In recent years the toy business in the US has become a twenty-three billion dollar industry. An ever increasing number of today's toys contain integrated circuits capable of storing and playing back audio.

I'd like to take you on a tour of the application of digital audio in Toys. We'll start with a brief history of electronic toys, continue with a thorough discussion of today's available technology and finish up with some thoughts on the future of electronic toys. Afterward, I'll try to answer any questions you may have on the subject.

[1. A brief history of talking toys.](#)

The first toy to incorporate electronically synthesized speech was Texas Instruments' "Speak 'n Spell", released in 1978. Texas Instruments went on to release a number of educational games based on their speech synthesis chips and became a major supplier to the toy industry.

In 1984 a startup company, ESS, was formed to provide single-chip audio IC solutions. These devices were based on a patented time-domain speech compression technique which had its roots in the theories of linguistics. Compressed audio files were hand-assembled on a personal computer and later de-compressed by the audio IC during playback. In essence, parts of syllables of words were identified and repeated sounds were only stored once, to be re-used as required. This process can produce very natural speech at very low data rates, and still sees limited use today. As the compression process is quite time-consuming it has, for the most part, been replaced by more automatic methods. Products to use this process included Early Talking Books by Western Publishing.

During the 80s, Japanese companies like OKI, Epson and then Korean companies like Samsung developed systems based on 4-bit CPUs. They produced high quality ADPCM speech and could drive LCD displays however they were too expensive for use in toys.. ADPCM is probably the most popular method of speech compression in use today. I will define this later.

In 1987, another startup company, ISD (Integrated Storage Devices) began developing a unique record/playback technology called ChipCorder. Developing a previously unused characteristic of EEPROM memory, which allows them to store different voltage levels, not just ones and zeros, in each memory cell and by sampling and storing the audio signal directly into these memory cells without converting them to digital values, ISD was able to make high quality audio recordings on a single chip. In this way, these devices can store an audio signal with eight times as much resolution as a similarly sized digital storage device. ISD's products have seen widespread use in our industry with such applications as talking greeting cards and the popular "Yak Bak".

Enter the Taiwanese Fab houses (or Integrated Circuit Foundries). Originally set up to service the expanding personal computer industry and hungry for more business, these Fab houses soon found a viable business in inexpensive melody chips - the type we have all seen in greeting cards, etc.

The first significant application an eight-bit microprocessor in a toy was developed in Taiwan in the early 1990's. Techno Mind was involved in the development of an 8-bit CPU based toy as early as 1992. In 1996, Techno Mind was involved in the development of an 8-bit CPU based toy that sold more than a million units! The most common 8-bit CPU used in the toy industry is nearly identical to the one that powered the famous Apple II personal computer! Today, there are multiple sources for these chips, and prices continue to drop.

2. A discussion of the current state of the art in electronics for talking toys.

- Most electronic toys are "systems on a chip"

What's available today:

First a few words about cost. The cost of an integrated circuit chip is dependent on two major factors: development costs including licensing; and die size. Generally, in our field, development and licensing costs (associated with the chip itself) are low - though there are exceptions. This leaves us with die size. As fabrication processes improve, more and more circuitry can be placed on a smaller and smaller die. This is the real reason why the ICs we use are so affordable. Prices will continue to drop and, as this happens, these ICs will find still wider use. Most of the ICs that I will now describe cost between \$0.20 and \$2.00. Usually, these chips can also play back simple

melody using one or more simultaneous (square wave) tones such as those you might find on a greeting card or novelty item..

Nearly all of the products listed below employ Masked ROM, which is used to store all recorded audio and firmware. By its nature, Masked ROM implies that we are working with semi-custom integrated circuits. "Standard" chips are built up with all but the last few layers and placed in stock by the FAB house. When an order is received, the final layers, which include the Masked ROM, are added.

- [Chip-on-board \(COB\) is a process in which an IC is bonded directly to the printed-circuit board.](#)

For inexpensive ICs, approximately fifty percent of the cost is in the package. Chip-on-board, or COB, is a way to reduce this cost. Instead of being separately packaged and treated as an assembly component (along with resistors, capacitors and the like) the IC is bonded directly to the printed circuit board. Essentially, the chip becomes part of the printed circuit board.

IC chips are delivered in "dice" form. Each die is somewhere between 1/8" and 1/4" on a side and less than one mm thick. After they have been fabricated the dice are placed in special containers one-hundred at a time. Bonding these chips to the printed- requires an investment in specialized "bonding" equipment as well as clean rooms and special handling equipment. In the bonding process tiny aluminum wires are welded to the pads on the die and directly to traces on the printed circuit board. An operator watches a TV monitor which displays a greatly magnified view of the surface of the IC. The operator must manually line up the target to two points on the IC. Once this is done, the process is automatic. Wires are welded from all relevant pads on the IC to the corresponding traces on the pc-board. When this process is complete the board is placed in a test fixture. Units that pass are then hermetically sealed using a blob of epoxy.

- [Voice ROM, 4-bit and 8-bit embedded controllers \(today's average toy is more powerful than an Apple II computer\)](#)

Voice ROM

This is one of the older technologies used in electronic toys. It is also seeing less and less use as much more flexible solutions become more affordable. Voice ROM chips contain the recorded audio plus the hard-wired logic circuitry necessary to play it back. A limited number of options, such as the number of different "triggers" and how they respond are available at design time.

Pricing is generally quite competitive, however. As a result of their low price, minimum order quantities are high.

Applications for Voice ROM chips are extensive and varied. Anything from greeting cards to promotional materials to simple non-interactive toys can be designed using Voice ROM. These are products which require a specific non-varying response to a specific stimulus.

ISD's ChipCorder Record/Playback Technology

In the late 1980s Integrated Storage Devices introduced a unique new technology based on the EEPROM (Electrically Erasable Programmable Read-Only Memory). Instead of only ones and zeros, two-hundred and fifty-six different voltage levels are stored. Audio signals are sampled and stored, without being converted to digital values. In this way the effective storage capacity dramatically increased. For a given amount of silicon real estate four times as much data can be stored on a ChipCorder compared with a common compressed digital audio format known as 4-bit ADPCM. Sampling rate is still an issue, however the chip contains the necessary input and output filter circuitry to reduce the effects of aliasing to an acceptable level. Sound quality, while limited in bandwidth, is quite good. Background noise is relatively high, however it sounds like analog noise and does not have the hard grating quality we normally associate with small-wordsize digital audio. ISD claims a one-hundred year data retention capability and 100K record cycles. In all likelihood the chip will outlive all of its support electronics. Perhaps archeologists of the future will mine our landfills for snippets of speech recorded by today's children!

ISD's ChipCorder technology is currently available in various sizes ranging from ten seconds to sixteen minutes. To my knowledge, this technology cannot be integrated with a CPU on the same chip, though two-chip solutions are common. A recent new addition to their line is capable of storing both analog and digital information on the same chip. This opens up a whole new set of possibilities as far as applications are concerned.

Four-Bit "Tiny" Controller plus Voice ROM

I briefly mentioned 4-bit microcontroller-based products when I spoke about toy history. These devices contain a limited amount of I/O (for example: three 4-bit ports – one read-only, one read-write and one write-only) and one or two analog outputs capable of driving a single transistor audio amplifier. A built-in clock-oscillator is implemented and generally operates using only an

external resistor to set its speed. A crystal or external clock can be used, however these are uncommon in an environment where every penny has to be watched.

Four-bit CPU based chips represent a significant advance in flexibility over Voice ROM chips. Because they execute user instructions they can be programmed to respond differently to various situations. Small programs are easily implemented. Simple interactive toys and games are easily designed using these chips.

What are the limitations? There are two things that four-bit tiny controller chips cannot do. 1) There are no interrupts. With the exception of a “wake-up-from-sleep” function, ports must be polled constantly while the unit is operating. As a result they have difficulty doing two things at once (or multitasking). 2) There is no ALU (Arithmetic Logic Unit), so mathematical operations are difficult if not impossible. Counters and basic score keeping can be implemented (by manipulating registers), but that’s about it.

I should also mention at this time that with the popularity of Infra-Red (IR) remote control in toys the chip manufacturers have begun adding the required driver circuitry to their chips. This results in substantial reductions in cost and parts count. Thirty-eight kHz carrier systems are fully implemented, resulting in reliable infrared communication. This feature is now available in some of the four-bit and many of the eight-bit chips that are available.

Eight-Bit CPU plus Voice ROM

Most of the 8-bit CPUs found in toys are based on the venerable 6502 chip. These can be complete (licensed) or partial implementations of these chips. (Western Design Center owns the rights to this architecture and actively maintains it.) Clock speeds of 4 MHz are common.

Chips based on 8-bit CPUs are available in a number of configurations. Some have LCD driver circuitry included while others have a built-in 8-bit A/D converter and SRAM enabling record/playback functions. It should be noted at this time that the Apple II line of personal computers was based on a much slower version of the 6502 chip, as were the Commodore and Atari computers. As far as toys and games are concerned, the limit on what these chips can do is the imagination of the designer and programmer.

- [On-Chip D/A converters](#)

There are two types of D-A converters used in toy-grade microprocessors. Both are common. Some chips implement both types simultaneously. The first, the Pulse Width Modulator, is common where no additional amplifier is available though it can just as easily be used to drive either a discrete or packaged amplifier. PWM outputs are usually configured to drive a speaker in bridge mode for extended output capability. The second type, current drive D-As are capable of driving a single transistor class-A amplifier with a minimum external parts count (one resistor). The output is fed directly to the base of the transistor. Information on the actual implementation of these converters is scarce.

- Audio amplifiers

Examples of amplifier circuits are common in data sheets and applications notes, and vary according to the type of D-A used. When an external amplifier is used, it is usually a single transistor class-A circuit. The loudspeaker is connected directly between the battery and the collector of the transistor. While this results in high current consumption and significant voice-coil heating, it has been used successfully in millions of toys! Care must be taken to turn-off the D-A converter when no signal is present so that there is no unnecessary drain on the batteries – not to mention heating of components. This is usually accomplished by “playing” a specially constructed sound file that ramps the output current down to zero. Conversely, another file is used to (silently) ramp the D-A output up to $V_{cc}/2$ before playing a sound.

Since PWM outputs are usually bridging, it is common to see a four-transistor “H-bridge” amplifier.

One design called “Green Voice” was originally developed by Peter Lam while he was at Mattel. It was intended to be an industry standard and can increase battery life as much as 100 percent over conventional circuitry. Green Voice uses both PWM and Current Mode outputs simultaneously in an effort to maximize output level while at the same time making dramatic improvements in battery life. A new series by Holtek features Green Voice technology, incorporating most of the circuit into the IC chip. Only one external transistor is required.

Where output level must be high or audio quality better one can use an IC power amp such as SGS Thompson’s TDA-2822, National’s LM4862 or TI’s TPA301. These devices have been optimized to operate at low voltage, however they are expensive by toy standards and see only occasional use. Some incorporate internal shut-down circuitry to save on battery power when no signal is present. Otherwise, an external switching transistor must be used to turn off the

amplifier. Either way, an output bit on the CPU must be used to control the amplifier. "Play" routines in software must turn the amplifier on and off each time data is sent to the D-A.

- Reconstruction Filters

Due to cost considerations, little or no attention is paid to the use of reconstruction filters in toys. Usually it is left to the mechanical characteristics of the loudspeaker. If necessary, a first order filter is added. This can take the form of a capacitor placed across the speaker or a series R-C network between the collector and base of the amplifier. If an amplifier IC is used, it is often possible to implement a second- or third-order filter (in a Sallen & Key configuration), however I have not seen this done. Again, cost is the controlling factor.

- Loudspeakers

Loudspeakers are chosen based on price and sound quality. Most have either paper or mylar cones and vary between 3/4" and 3" in diameter.

- Enclosures?

I have yet to see any examples of porting or other such attention paid to enclosure design in toys. At this point in time the best we can hope for is that the enclosure doesn't rattle too much!

- Speech Recognition on a Chip

This is a technology that is beginning to see widespread use, as witnessed by the number of product offerings at recent Toy Fairs in New York. Basically, there are two kinds of speech recognition: speaker dependent and speaker independent. It is important to understand the difference.

Speaker dependent speech recognition is "tuned" to the characteristics of one person's voice through a training process. It is more accurate (99%) than speaker independent recognition (96%), but can only recognize words spoken by a single user. Speaker dependent speech recognition is also becoming popular in the world of Personal computers. You may be familiar with IBM's "Via Voice" or Dragon's "Naturally Speaking".

In the case of Speaker Independent speech recognition the unit can recognize anyone's voice (within reason). This is accomplished by training the unit to a broad sample of people's voices. In the United States, the sample must include (~500) people from all parts of the country to account for different dialects, etc. Examples of speaker-independent speech-recognition systems are beginning to see widespread use by telephone companies.

Chip-based speech recognition has a much smaller vocabulary than its PC-based counterpart. While it is possible to store a larger number of words, it is common for these devices to only be able to work with twenty words at a time. Newer chips have word-spotting capability, wherein words can be picked out of phrases. Previously, the speaker had to be cued as to when to say a word, severely limiting functionality. Designers and programmers have to be careful not to allow situations where similar-sounding words are expected. Also, much attention must be paid to prompting the user to use the correct set of words.

The Sensory RSC series of Speech Recognition chips incorporate a microphone pre-amplifier, A-D, D-A converters, a complete 8-bit RISC microprocessor and (optionally) 64 Kbytes of masked ROM for firmware and SI recognition weight sets, all on one chip. External flash memory is required for Speaker Dependent recognition. Speech recognition is accomplished using a Neural Network. Their current top-of-the-line chip, the RSC-364 also incorporates a 24 x 24 hardware multiplier and additional (scratch pad) RAM to reduce calculation times as well as licensed noise-reduction technology from Sarnoff Corporation.

Besides doing speech recognition, the Sensory chips can record and play back audio using an external flash memory chip.

Another chip-based speech recognition system was recently unveiled by ISD. To my knowledge it is not priced to be competitive in the toy arena.

Music Processor

This type of chip is capable of generating high-quality music similar to what you might expect from a PC sound card. In fact, these chips contain many of the subsystems contained in a sound card. They can generate sound in three ways: 1) synthesized waveforms, 2) sampled sounds, 3) sampled waveforms - in a manner similar to the popular "wavetable" synthesis. Since melody information is stored as command sequences - somewhat akin to midi files - long durations are possible. Due to the memory requirements of the sampled sounds and waveforms these chips tend to be moderately priced.

Record-Playback Processor

These chips feature an on-chip Analog-to-Digital converter and SRAM and are thus capable of recording and playing back small amounts of audio and other signals. They are capable of performing limited DSP (Digital Signal Processing) on the recorded signal such as pitch changing, echo and other effects.

3. Digital Audio for Toys

- Sampling Rate and Word Size

One measure of the accuracy of a Digital Signal is its word size. A Compact Disk uses a 16-bit word, a sign bit plus 15 value bits, allowing 65,536 unique values and a best case dynamic range of 96 dB.

Anyone who has played with an electronic toy or game will tell you that the sound is not CD quality. Why is this? Let's do some simple calculations. CD-quality audio has a Sampling Rate of 44,100 samples per second and a word size of 16-bits (or two Bytes). For a monaural signal each second of recorded material consumes 88,200 Bytes of storage. ROM memory is usually calculated in Bits - Marketing people always use the largest number. So one second of CD-quality audio consumes 705,600 bits. To put this in perspective, the largest Voice ROM chips are now in the 8 Megabit range, enough to store about 11 seconds of audio. Considering that such a chip costs a few dollars, this simply isn't practical.

If we lower the word size to 8-bits, and compress it further to 5-bits (4-bit ADPCM is also common, though sound quality is significantly worse), and choose a more practical sampling rate of 6,000Hz: Let's see... the same 8 Megabit chip could store about 273 seconds of audio - about four and one half minutes! We all know that you don't get something for nothing, what did we lose when we increased our playback time by 24 times? Well, by changing the sampling rate to 6,000Hz we limited our bandwidth (or frequency response) from 20,000Hz to 3,000Hz! If we are trying to reproduce just speech, this is an acceptable choice. To put this in perspective, a telephone has a frequency response that is limited at the upper end around 3,000Hz. Few will argue that the telephone does not reproduce speech satisfactorily.

What else have we lost? A few moments ago we determined that CD-Quality audio had a theoretical dynamic range of 96dB. If we reduce the word size to 8-bits, what dynamic range are we left with? We said earlier that an eight-bit word could hold 256 unique values. Compare this to 65,536! In decibels, this translates into a theoretical dynamic range of 48 dB. Note that is very easy to calculate theoretical dynamic range as a function of word size. Each additional bit provides an additional 6 dB of dynamic range. I should also note that the theoretical maximum is never realizable. It is probably realistic to subtract 10 dB in each case. This is due, in part to noise and in part to the limitations of the conversion circuits.

Can we work within the constraints of a 38 dB dynamic range? The answer is yes, though we must pay careful attention to signal levels in the digitizing process.

What other problems do we encounter in the digitizing process. Well, one thing that we hear quite often is Quantization Noise. This is caused by the fact that the original signal had values that did not fall exactly on the 256 possible values of our 8-bit word size. The difference (error) varies from sample to sample and, as it turns out becomes a signal of its own - you guessed it - a noise signal! This type of noise is very apparent under two conditions: 1) The original signal is not a full-scale signal. Let's say that our signal is 10 dB below the maximum possible value. Now we are only 28 dB above the noise 2) Our original signal is a pure tone, or has relatively few harmonics. This guarantees that our Quantization Noise signal is as loud as it can be.

- Compression Schemes - ADPCM, LPC, CELP, MELP

LPC Linear Predictive Coding LPC was first developed in 1982 by Bishnu Atal at AT&T Bell Labs . (It is interesting to note that Bell Labs demonstrated a working Vocoder at the 1939 World's Fair!) LPC is based on a simple model of the human vocal tract which consisted of two signal generators a periodic waveform generator (for "voiced" sounds) and a white noise generator followed by a filter (which loosely modeled the rest of the vocal tract). Encoding required a lot of computing power and usually the talents of a PHD linguist. The various parameters were adjusted iteratively until error was minimized. The stored data consisted of controls for the various elements of the model: Source, pitch, envelope and filter coefficients. Speech quality was poor and generally considered to be quite robotic sounding.

CELP Code Excited Linear Prediction CELP takes the concept of LPC to another level by considering the error signal which had been abandoned in the LPC process. As the error signal contains just as much information as the original speech, it has to be compressed also. In CELP,

the error signal is compared to a “codebook” of anticipated error signals and a best match found. Now only the reference to the codebook entry had to be stored. CELP provides a significant improvement in speech quality at a cost of about two times the data rate of LPC.

ADPCM Adaptive Delta Pulse Code Modulation This is more a compression scheme than a data format, but I guess you could say it is both. Here, instead of storing the actual value of the current sample, what is stored is the difference between a predicted value for the current sample and the actual value. ADPCM usually results in a reduction in word size of three to four bits. Generally, ADPCM implementations are table-based. In order to reduce the load on the CPU a lookup-table is implemented in code and used in lieu of direct multiplication of data.

MELP

Texas Instruments is re-entering the toy arena with a new technology called MELP and a family of chips that incorporate it. These chips have two significant advantages over previous ones. First, they allow a significant increase in the amount of speech that can be stored on a given chip while at the same time providing good audio quality. Second, and perhaps more important from the perspective of previous TI designs, the process of encoding audio data for the chip is greatly simplified. Additionally, these chips allow the user to change formats on the fly during playback. So a segment that includes speech and music can be divided up into segments according to the type of encoding that best suits them (PCM, ADPCM or MELP). This results in the highest quality sound with the best possible compression.

[- Post-processing audio for toys \(breaking all the rules\)](#)

There are two common approaches to digitizing audio for use in toys. Until recently, the most common approach was to use proprietary hardware and software that was provided by the Fab house that made the chip that you were using. This approach had a unique set of strengths and weaknesses. Interestingly, its major strength lay in the fact that it used an 8-bit analog-to-digital converter. Provided you had a clean audio signal to work with, it was relatively easy to get good results. All you had to do was to be willing to drive the A-D converter fairly hard into clipping. For reasons that I do not fully understand, the distortion due to clipping appears to be much less objectionable than that of quantization noise! First and foremost among the weaknesses of this approach was the editing software that came with these units. Usually an MS-DOS -based program, it was hobbled by arcane commands and limited editing capability. Once a selection was digitized, all you could do with it was simple cut, paste, trim and mute operations. There had to be a better way.

When I started at Techno Mind I brought with me a shareware wave file editor called Goldwave. I had been introduced to Goldwave through Syn-Aud-Con, where it was popular for its function generation ability. You can mathematically define a waveform and Goldwave will generate it for you! Being familiar with its ease of use, I immediately tried to use it as a substitute for the proprietary systems. To my immediate surprise, it was very difficult to get the kind of results with Goldwave that my peers had come to expect using the proprietary systems. Yes, it was much easier to work with. Large segments of recorded material could be digitized in a short time but the quality of the sound when put on a chip was inferior. What was wrong?

An analysis of the problem led to some interesting results. It turns out that the biggest problem facing the toy sound engineer is not sampling rate but word size. If you take a 16-bit wave file, realizing that you probably have 15 bits or so of resolution (losing one bit to conversion error), that's about 90 dB of dynamic range. If quantization noise robs us of another bit, we theoretically have an 84 dB signal-to-noise ratio. Now let's consider 8-bit audio. Seven bits of resolution gives us 42 dB of dynamic range. Quantization noise is on the order of -36 dB in theory. In practice it appears to be worse than that.

OK, so let's try normalizing, EQ-ing, compressing and noise-gating the signal (not necessarily in that order). This helps a lot, but now we are doing a lot more work - especially when you consider the fact that due to differences in the program material each word or phrase has to be handled separately if you want to achieve optimum results.

In order to make this practical we have to compromise. We use a batch converter such as Wave Convert by Waves or the Batch Converter Plug-In for Sound Forge. The batch converter processes all of our files using a pre-defined sequence of operations. Afterward, we must listen to every file and do further work on those files that still need attention. A typical medium size job requires three days of recording and two to four weeks of post production!

[- Other Issues - using the CPU and a vocabulary of recorded words to generate speech.](#)

In order to get the large quantities of speech claimed in many toys, it is necessary to record a vocabulary of (200 - 300) words and have the CPU string them into sentences or phrases. In theory, this should be relatively easy. Certainly, it is easy to write assembly language routines to choose and play back the required words. Unfortunately, this process turns out to be quite difficult.

First, if you record each word individually and try stringing the words together the result sounds extremely robotic. The next obvious method is to try to capture individual words out of recorded phrases. Sometimes this is pretty easy, but usually, it becomes extremely difficult. It turns out that people start saying the next word before they finish saying the current word. All the words become irrevocably intertwined (or polluted) with adjacent words. Reasonable results can be had if one records a mix of individual words and phrase segments, tries to pick words out of phrases where possible, then picks the best from each group and carefully re-assembles every sentence to make sure that everything sounds good. This is how it is done. It is very time consuming, but until a better way comes along it is the preferred method.

Until recently, Text-to-Speech made too many demands of a toy grade CPU (in terms of processor speed and memory) to see any use. This may soon change. This year Mattel introduced a robot toy that features text-to-speech capability. Note that the robotic sound of text-to-speech is used here to advantage.

[4. What will the future bring](#)

MPEG on a Chip

Low cost MP3 decoder chips are just now coming on line. These chips promise near-CD quality audio for a variety of portable consumer products. Presently they are still too expensive to see widespread use in toys, but not for long.

Multi-Chip Solutions

As chip prices continue to fall, and die sizes reach their practical limits (at least in terms of the current foundries) it is becoming cost effective to offer two-chip solutions. These are often divided into CPU and Masked-ROM units.

Already, Speech Decompression companion chips based on CELP are available that can increase the capacity over that of a single-chip solution by six to ten times for a cost of about one dollar. At present this is only satisfactory for speech, but we can expect to see improvements. Applications include talking books.

DSP solutions are also on the horizon. These devices will be popular because they will provide significant advances in both sound quality and chip duration. Eventually, one can expect the sound quality in toys to approach that of a portable radio, for instance: