Designing in Text-To-Speech Capability for Portable Devices

Market Dynamics

Currently, the automotive, wireless and cellular markets are experiencing increased demand from customers for a simplified and more natural user interfaces. This is because cellular phone keyboard and screens are tiny, difficult to use and have bulky menus. Designing in Text-To-Speech capability for handheld devices will make consumers’ lives easier and more productive. For example, enabling users to safely listen to emails downloaded to the device while driving.

Telematics and infotainment have become popular in the automotive world and speech-enabled car accessories are designed to help improve safety, leaving the driver with the most important task of all - driving the car. A recent study conducted by the National Safety Council suggests that conversing on cell phone while driving can lead to significant decreases in driving performance. The study found that driver distractions due to cell phones could occur regardless of whether hand-held or hands-free cell phones are used, and that cell phone conversations create much higher levels of driver distraction than listening to the radio or audio books. A combination of self-awareness, as well as legislation, is driving the trend towards hands-free speech-enabled devices in the automobile.

Listening to customized text that has been converted to speech (utilizing embedded text-to-speech solutions) will become more common as consumers begin to realize that reviewing the latest stock quotes, news, sports scores, etc., using the handheld devices saves time otherwise wasted while driving. There also exists a growing demand for ease of use GPS systems that verbally give directions and eliminate the need for a screen. Finally, most consumer devices come with lengthy manuals that require time and study in order to master. Text-to-speech enabled devices can all but eliminate the burden of these manuals by providing users with verbal prompts and audio directions for operation.

Text-To-Speech technology can also provide crucial services to people with certain disabilities such as the visually impaired, speech or hearing impaired people, and dyslexia – thus helping them communicate. And improving life quality and simplicity is what we all seek, isn’t it?

In summary, there exist many applications where speech synthesis could be useful such as: automotive, GPS, telephony Caller ID, learning aids, toys, e-books, measurement and industrial systems.
Machines That Speak Like a Human – The Challenge

Several text-to-speech synthesis approaches are currently available on the market, such as rule-based-synthesizers, articulatory-synthesizers and concatenative synthesizers. The rule-based-synthesizers try to describe speech elements by parameters related to formant frequencies, bandwidths and voicing. Articulatory-synthesizers try to imitate the physical human mouth, wherein each speech element is described by parameters of the actual human mouth’s position and movement. Concatenative-synthesizers consist of a speech elements corpus that is excised from an actual speech recording, with linguistic rules to select the units accordingly and concatenate them to produce speech.

Each approach differs from the other and offers reasonable quality. However, the concatenative approach seems to be the industry’s current favorite as it can offer the most natural sounding speech. The disadvantage of using this approach is the tremendous amount of memory required for the implementation, thus making it an expensive solution for the embedded environment.

High quality speech synthesis at an affordable price is difficult to find in today’s market. Furthermore, most high quality text-to-speech solutions require huge amount of memory and powerful embedded processors to accommodate the synthesis mechanism. These solutions often require a server or PC that can provide almost unlimited resources. Developers wishing to add speech synthesis functionality to existing products currently would have to use additional expensive components (memory, processor and analog circuits), port the Text-To-Speech software package to the environment and be knowledgeable in speech synthesis algorithms.

Integrated, Single-Chip Solution

Given current cost and technology restrictions, single-chip, integrated solutions are the next step. Winbond provides one such solution, a Text-To-Speech processor (WTS701) using concatenative synthesis approach. The IC accepts ASCII/UNICODE format input via a serial SPI port and converts it to spoken audio via an analog output, directly driving a speaker or a digital CODEC output.

Figure A. Winbond’s Text-To-Speech General Operation.
The WTS701 integrates a text processor, smoothing filter and multi-level storage array functions on to a single-chip. Text-to-speech conversion is achieved first by analyzing the incoming text to normalize common abbreviations and numbers into a spoken form. The normalized text is then analyzed for phonetic interpretation, with the translation mapped into samples to be played out of the analog storage array. This output signal is then smoothed by a low pass filter and is available as an analog signal, or it can be passed through the CODEC for digital audio output.

![Figure B. The WTS701 in Analog and Digital (CODEC) Environment.](image)

The synthesis algorithm attempts to use the largest possible word unit in the appropriate context thereby maximizing natural sounding speech quality. The speech units are stored, uncompressed, in a multi-level, non-volatile analog storage array to provide the highest sound quality to memory density trade-off. This solution is made possible through Winbond’s patented multilevel storage technology. Voice and audio signals are stored directly into solid-state memory in their natural, uncompressed form, providing more natural sounding voice reproduction.
Changing languages used to provide a number of technical challenges but the WTS701 can be programmed through the SPI port, allowing download of different language and speaker databases. Currently supported languages include U.S. English and Mandarin, with other languages in development or planning.

To make integration easier, Winbond implemented a user-friendly interface with many useful commands to control the chip, and in addition provides the option of changing the abbreviations by adding or deleting abbreviations according to the customer’s preferences. The chip also features low power consumption, power down mode support and real time conversion for streaming text, making it ideal for embedded systems applications.

The WTS701 is based on Winbond’s Multi-Level Storage (MLS) technique in which one of 256 distinct voltage levels are precisely stored per memory cell. This provides eight times more storage space for any given memory size than the ordinary digital signal storage technology which can store only 0 or 1.

Since Winbond implemented the Text-To-Speech processor using its Multi-Level Storage technology, it can actually store real human voice recordings, thus contributing to superior natural human sound. This audio quality is unprecedented when compared to other Text-To-Speech synthesizers currently available in the market, which provide synthesized robotic sound and unpleasant to listen to. The MLS methodology makes Winbond’s solution an effective one, as the memory size required to store the same amount of audio samples in a conventional solution can be as much as 64 Mega-bit. Moreover, Winbond’s technology makes it possible to record a person’s voice and store...
it on the chip. These features could open up a myriad of popular consumer features, if for example, a device offered the voice of a popular celebrity speaking.
Figure D. Winbond’s Multi-Level Storage (MLS) technology.

Conventional Storage
Stores 1 of 2 voltage levels
1 “bit” of data per cell

Winbond Multilevel Storage
Stores 1 of 256 distinct voltage levels
8 equivalent “bits” per cell

Figure D. Winbond’s Multi-Level Storage (MLS) technology.
Winbond’s Text-To-Speech System Characteristics

As a real System-On-Chip solution, the WTS701 performs the overall control functions for both host controller and text-to-speech processing.

The WTS701 system architecture consists of the following functions:

- Serial interface to monitor the SPI port and interpret commands and data.
- Text normalization module to pre-process incoming text into pronounceable words.
- Words to phoneme translator, which converts incoming text to phoneme codes.
- Phoneme mapping module that maps incoming phonemes to words, sub-words, syllables or phonemes present in the MLS memory.
- Volume and speed adjustments.
- Digital and analog output blocks for off-chip usage.

The WTS701 system performs text-to-speech synthesis based on concatenative samples. The units for concatenation can vary from whole words, to syllables, and right down to phoneme units. The perception is that the larger the sub-word unit used for synthesis, the higher quality the speech output. A corpus of pre-recorded words is stored in Winbond’s patented MLS memory and a mapping of the various sub-word parts is held in a lookup table. The speech creation is achieved by concatenation of these speech elements to produce words. The system process flow is shown in Figure E.
Figure E. WTS701 System Process Flow.

The text to speech component of the system consists of three principal blocks:

- Text normalization
- Letter-to-phoneme conversion
- Phoneme mapping

Text Normalization

Text normalization involves the translation of incoming text into pronounceable words. This includes such functions as expanding abbreviations and translating numeric strings into spoken words. It involves a certain amount of context processing to determine the correct spoken form. In addition, the WTS701 has an abbreviation list stored in the device’s internal memory to convert all appearances of acronyms, abbreviations or special characters (such as SMS icons) into the appropriate text representation.

The default abbreviation list supported by the WTS701 is a general one and can be modified by the user, to match the domain that the text is being loaded from. This creates flexibility for adding abbreviations specifically for the text, either by the
developer or even the end user to best customize the product for its text application. SMS unique characters can be supported through this functionality as well, defining the icon ASCII/Unicode text and its replacement.

**Words-to-Phoneme Conversion**

Once the data stream has been translated to spoken words the system next determines how to pronounce them. This function is obviously highly language dependent. For a language such as English it is impossible to break this task down to a set of definitive rules, but easier for Mandarin (Beijing dialect) because Mandarin has a simpler tonal system, possessing only four tones. The task is achieved by a combination of rule based processing coupled together with exception processing.

**Speech-Element Mapping**

This algorithm maps phoneme strings into the MLS phonetic inventory. This task falls into two portions. First, the word must be split into sub-word portions. This splitting must be done at appropriate phonetic boundaries to achieve high quality concatenation. Once a sub-word unit is determined, the inventory is searched to determine if a match is present. A matching weight is assigned to each match depending upon how closely the phonetic context matches. Each sub-word has a left and right side context to match, as well as the phoneme string itself. If no suitable match is found in the inventory, the sub-word is further split in a tree-like manner until a match is found. The splitting tree is processed from left to right and each time a successful match occurs the address and duration of the match in the corpus is placed in a queue of phonetic parts to be played out the audio interface.

**Conclusion**

Text-To-Speech is a very promising technology, fulfilling a very real need in the rapidly growing communications market. Helping people communicate, promoting safety, increasing productivity, improving convenience, and conveying information in more user friendly ways are just some of the benefits provided by adding a text-to-speech feature. Winbond’s Text-To-Speech is the first real System-on-Chip solution, and a significant milestone in the effort to bring this feature to the embedded world. This solution reduces time-to-market, eliminates Text-To-Speech development and integration costs, and it is best for battery powered and space sensitive devices.

However, expectations must be set properly, since embedded systems that have low resources and limited space cannot provide the same quality as the PC and server based solutions. Thus, Winbond’s device is a step up, a move in the right direction, providing the simplest way to integrate this feature, and provide superior sounding speech.