CIRCUIT CELLAR ONLINE

THE MAGAZINE FOR COMPUTER APPLICATIONS

Circuit Cellar Online offers articles illustrating creative solutions and unique applications through complete projects, practical tutorials, and useful design techniques.

RESOURCE PAGES

RESOURCE LINKS

A Guide to online information about:

Speech Synthesis

by Bob Paddock

This month, my Resource Pages cover speech synthesis and speech recognition. In some cases, I could not make up my mind about which page something should be in because it seemed equally fitting to both subjects. Because of this, there is some overlap between the two.

I'll cover the basics of speech synthesis via the FAQs, then go to what you really want to know, "What chips can I put in my project?"

Because this is a Presidential Election Year, I'll close with links to TruthVSA: Voice Stress Analysis Freeware. Do you think you could embed TruthVsa in a DSP chip?

Two good places to learn more about what is happening with speech I/O are the University of Essex Department of Language and Linguistics/SPEECH GROUP and the comp.speech Frequently Asked Questions site.

The FAQ site provides a range of information on speech technology, including speech synthesis, speech recognition, speech coding, and related material. They have "over 500 hyperlinks to speech technology web sites, ftp servers, mailing lists, and newsgroups." Makes my life easy this month.

Phonetics and Theory of Speech Production: Speech processing and language technology contain many special concepts and terminology. To understand how different speech synthesis and analysis methods work, you must have some knowledge of speech production, articulatory phonetics, and some other related terminology. The basic theory of these topics is discussed briefly in this chapter. For more detailed information, see Fant (1970), Flanagan (1972), Witten (1982), O'Saughnessy (1987), or Kleijn et al (1998).

The release of Holtek's HT85*xxx* series of Green Voice devices marks an important step in the range of devices available for speech synthesizer and melody generator application areas.

The HT817D0 is a single chip LOG-PCM voice synthesizer LSI with 16.8-s voice capacity at a 6-kHz sampling rate. The chip, when triggered, drives a speaker through an external transistor with a current switch D/A converter output. Negligible current will be consumed in the standby state.

Information Storage Devices, Inc.

ISD is famous for their digital tape recorder style of chips.

Long before ISD was around, I designed a similar digital tape recorder. Find out more about it in the Voice Section of ASKUS.

Let us help keep your project on track or simplify your design decision. Put your tough technical questions in front of the ASKUS team.

Oki is a major supplier of dedicated speech synthesizers.

- Play-only series
- Record/Play series
- Low power speech amplifiers
- Serial voice registers/ROMs/flash memory
- Special speech functions
- Voice recognition

The MSM7630 is a multi-lingual speech control processor (SCP) with text-to-speech synthesis capability in six languages, including American, English, French, German, Spanish, and Japanese. The speech processor is an LSI device with an internal D/A converter. It is optimized for speech output applications, such as text-to-speech conversion. A PDF of this datasheet (651 KB) is available but is a real pain to see. The only browser that worked for me was Netscape 4.72. Opera did nothing, and IE4 downloaded useless glop. I had to fill in one of those silly web forms for each datasheet I wanted, too.

- Features
  - Parallel and serial interfaces
  - Single 3.3-V power supply
  - 5-V interface available
  - Internal 16-bit x 16-bit to 32-bit multiplier (2-clock data throughput)
  - 26-VAX MIPS performance at 40-MHz operation (when using ordinary ROM/SRAM)
    Package: 100-pin plastic QFP (QFP100-P-1420-0.65-BK)(Product name: MSM7630GS-BK)

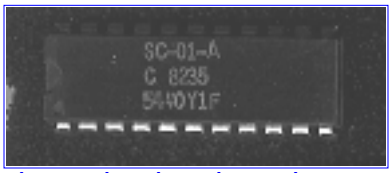
RC Systems has been a market leader of affordable, high-quality text-to-speech synthesis products since 1983. You'll find RC Systems synthesizers in a wide range of products, from talking sewing machines, to point-of-sale terminals and oil rig monitors, to space satellite telemetry systems.

RC Systems synthesizers are available as plug-in boards, modules, and chips. For PC-based applications, the DoubleTalk family is offered for PC, PC/104, and Apple platforms. The V8600A voice module and RC8650 chipset are ideal for use in embedded applications. Software licensing is also available.

For a bit of history on the original speech synthesizer, SC01/SSI263(A), I thought you might find the following of interest. I'll also cover many links to people trying to put some "feeling" into speech synthesizers today.

Red Cedar Electronics archived some of the data on the first monolithic phoneme synthesizer. They have a good bibliography on the subject, too.

SC-01A Speech Synthesizer



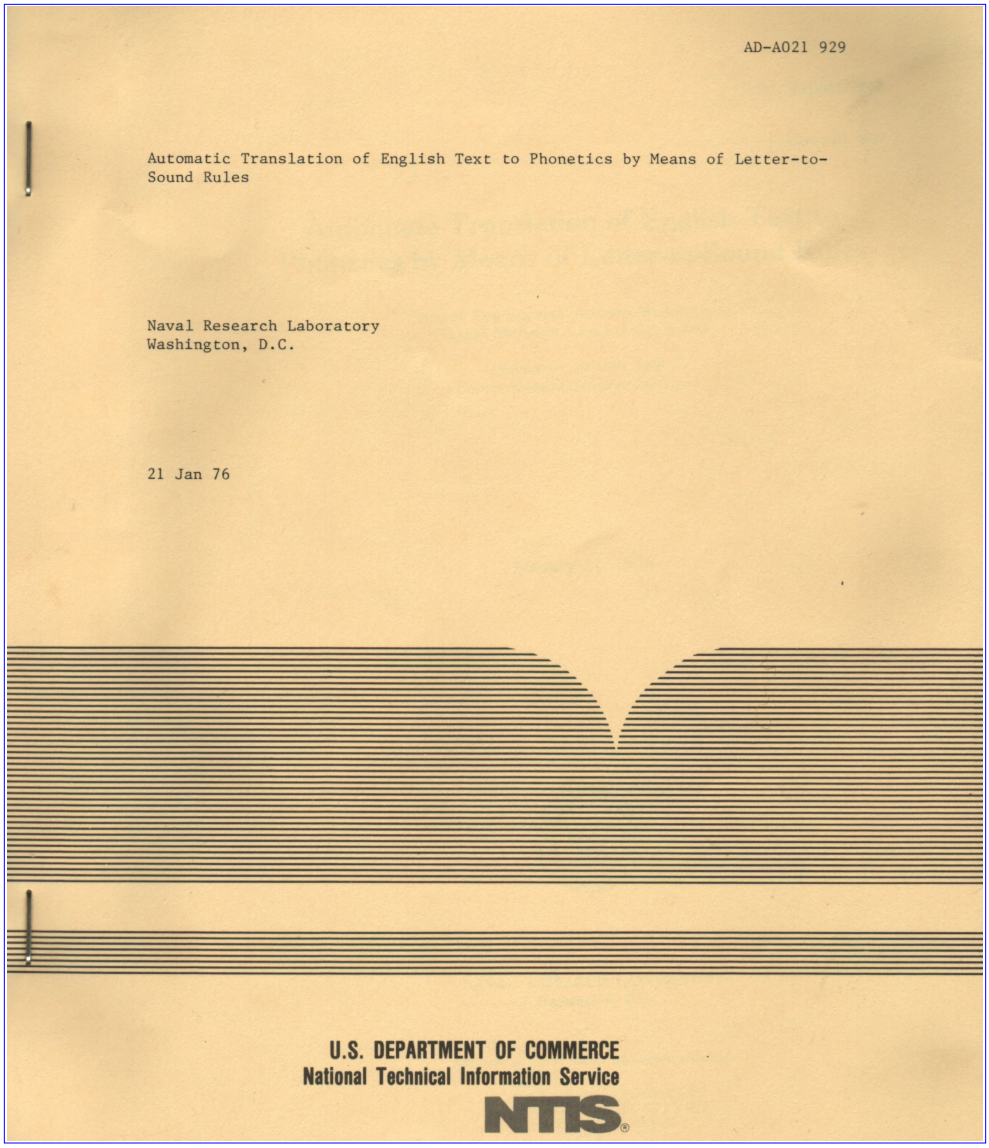
The Votrax SC-01A speech synthesizer is a phoneme synthesizer of the early 1980's.

For the theory of operation, see the following patents:

Voice Synthesizer, Mark Dorais (assigned Federal Screw Works (=Votrax)), 4,128,737,12/5/78

Voice Synthesizer, Carl Ostrowski (assigned Federal Screw Works), 4,130,730, 12/19/78

MITalk is described in "*From Text to Speech: The MITalk System*" by J. Allen, M.S. Hunnicutt, and D.H. Klatt, Cambridge University Press, New York, 1987.


AD-A021 929

Automatic Translation of English Text to Phonetics by Means of Letter-to-Sound Rules

Naval Research Laboratory
Washington, D.C.

21 Jan 76

U.S. DEPARTMENT OF COMMERCE
National Technical Information Service
NTIS

"Automatic Translation of English Text to Phonetics by Means of Letter-to-Sound Rules"
by the Naval Research Laboratory, Washington, DC, 1/21/76. It is document number AD-A021 929 from http://www.ntis.gov/.

I heard the Department of Commerce Secretary William M. Daley say he wanted to close the National Technical Information Service (NTIS) because "you could get everything on the Internet now." He is obviously as clueless as government officials usually are about technology, because you can't find many of the obscure papers on the Internet that you can find at NTIS.

U.C. Berkeley EECS225d Home Page Audio Signal Processing in Humans and Machines gives a history of text-to-speech from 1939 to 1985 in *Klatt Audio Scribe Notes for EE225d.*

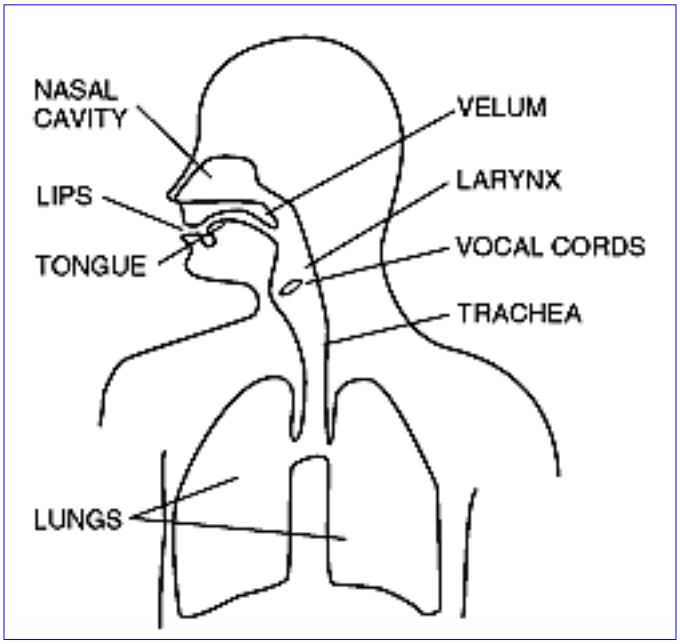
The next system after or in parallel with MITalk, depending on who's timeline you go by, is DECTalk, originally by Digital Equipment.

Work has continued to this day on improving DECTalk by third parties. Officially,

DECTalk is part of Compaq.

They say you can download a 60-day trial directly from them, but after a couple of hours of trying, I gave up. The site kept going in circles. It is Compaq product code ETD08-AA, if you want to share the agony of the search.

"*DECTalk Software: Text-to-Speech Technology and Implementation*" by William I. Hallahan is must read for anyone interested in text-to-speech. It covers how human speech is produced by the vocal cords in the larynx, trachea, nasal cavity, oral cavity, tongue, and lips. The figure below shows the human speech organs.



The text-to-speech synthesis assistive products information home page was created to provide a focal point of publicly available information about text-to-speech synthesis products. It is in no way officially associated with or sanctioned by any corporate entity. It makes an excellent starting place for learning more about how DECTalk is used.

With the background of DECTalk under your belt, the work of Janet E. Cahn directly answers the question of how to give your products words some "feeling."

Expressive Synthesized Speech Thesis, Cahn, Janet E., *Generating Expression in Synthesized Speech*, Master's Thesis, Massachusetts Institute of Technology, May, 1989.

Cahn, Janet E., *The Generation of Affect in Synthesized Speech*, Journal of the American Voice I/O Society, Volume 8, July, 1990, 1–19.

From this page you can hear the output of the Affect Editor program, which generates instructions for a DECtalk3 speech synthesizer.

Synthetic speech systems have not improved significantly in their ease of audition or their ability to express human-like emotion since the early sixties. To begin

addressing this problem, a prototype of a learning speech interface agent called TurnStyles has been designed and built. This interface agent dynamically learns critical pacing aspects of conversational style from "listening" to conversations and adapts the system's synthetic speech output to reflect the stylistic preferences of the user.

In a sense, TurnStyles enables a speech I/O system to "speak as it is spoken to," giving your device some feeling.

V*Star overcame the lack of feeling in DECTalk with their work in Avatars. It reminds me of the Sci-Fi movie "Looker" as to where this is all headed.

V*Star voices are optimized for graphical input of subtle meaning. Whereas, all conventional text-to-speech systems such as DECTalk are fully automatic and generally do not understand what is being said so they use a flat neutral inflection and intonation system (i.e., monotone). In contrast, V*Star's vocal editing technology allows authors to specify rich and varied intonation, inflection, and timing.

The aim of the Speech Synthesis Systems page is to present a cross-section of various speech synthesis systems. Some of these are academic, others are commercial. They represent many different techniques and will hopefully give you more ideas of what is currently possible with speech synthesis technology.

TreeTalk is a word pronunciation demo with (WAV or AU) speech output. The demo was developed by Bertjan Busser for his PhD project. It contains TreeTalk systems (IGTree decision trees trained on word-pronunciation pairs, performing both grapheme-phoneme conversion and stress assignment) for English and Dutch. More can be found on Antal van den Bosch's web page.

Sami Lemmetty did his Master's Thesis on the *Review of Speech Synthesis Technology* and has a excellent biography, *Speech Synthesis Literature.*

In 1996, Uwe Steinmann wrote *An Overview of Text-to-Speech Converter* that still gives a good, quick introduction to the subject.

The Centre for Speech Technology Research constantly updates a page of speech related links that is better than anything I could do on the subject. They try to cover all of the known work from all over the world.

You can download *Festival,* an extensible multi-lingual speech syntheses system, and the *Edinburgh Speech Tools Library*, a C++ library providing support for speech

processing, as well as more useful C++ classes such as containers and I/O utilities.

Try it yourself. It offers a full text-to-speech system with various APIs, as well an environment for development and research of speech synthesis techniques. It is written in C++ with a Scheme-based command interpreter for general control.

You can compare several speech synthesizers at the LDC / COCOSDA interactive speech synthesizer comparison site.

This site allows you to do side-by-side comparisons between text-to-speech systems and decide which one you prefer.

Select useful test text from a wealth of text corpora made available by the Linguistic Data Consortium.

The Department of Speech Music and Hearing at the Royal Institute of Technology covers several areas:

- Speech Communication & Technology
- Speech Signal Processing
- Centre for Speech Technology
- Music Acoustics
- Voice Research Centre
- Hearing Technology

The Royal Society for the Blind of South Australia Adaptive Technology Centre web site covers screen readers, synthesizers, talking applications, and voice recognition. You can download many different speech demo programs from their site.

Mission: To make spoken language systems work.

Get the speech toolkit and language resources and check out the *Survey of the State of the Art in Human Language Technology.*

The overall objective of the Speech Communication Group of the Research Laboratory of Electronics is to gain an understanding of the processes whereby (1) a speaker transforms a discrete linguistic representation of an utterance into an acoustic signal, and (2) a listener decodes the acoustic signal to retrieve the linguistic representation. The research includes development of models for speech production, speech perception, and lexical access, as well as studies of impaired speech communication. Check out MIT's Speech Communication Group.

Although nothing about speech is mentioned, I thought Atom Amplification (a new technique demonstrated by RLE researchers) was interesting .

You don't have to be a Star Trek fan to know that the computer of the future will talk, listen, and understand. That computer of the future is the Apple Macintosh of today. Apple's Speech Recognition and Speech Synthesis Technologies now give speech-savvy applications the power to carry out your voice commands and even speak back to you in plain English and Spanish.

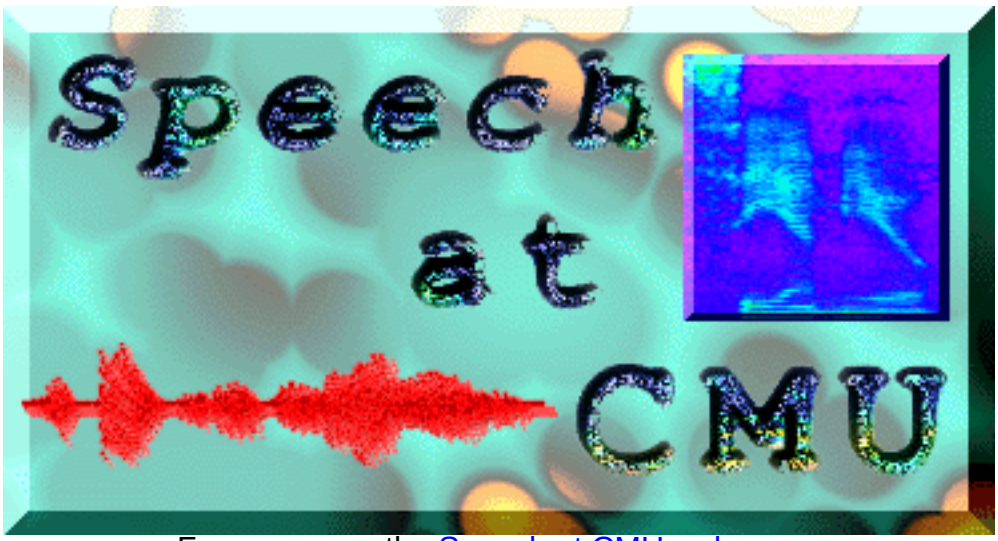
---

t2p: Text-to-Phoneme Converter Builder
Kevin Lenzo, Carnegie Mellon University.

t2p is a public domain package in Perl for building grapheme-to-phoneme rules from pronunciation dictionaries. In other words, it builds letter-to-sound rules for pronouncing words. It is given a set of example pronunciations, like from the CMU Pronouncing Dictionary. The Carnegie Mellon University Pronouncing Dictionary is a machine-readable pronunciation dictionary for North American English that contains over 100,000 words and their transcriptions. The CMU dictionary V.0.6, is freely available by anonymous FTP.

What would you use it for?

Because it can generalize words outside of the training set, it can be used to find the pronunciations of words the program has never seen.



For more, see the Speech at CMU web page.

Jason Woodard, Department of Electronics & Computer Science at the University of Southampton, does a good job of covering speech coding. He gives an idea of the principles involved in speech coding and details commonly used coders. Also, links are given to other related pages and the source code of some common speech codecs.

Low Rate Speech Coding by Clare Brooks, describing two speech coders operating at 1.9 kbps and 2.4 kbps, is worth a look at if you need to squeeze 10 MB of speech data into a 1 MB EPROM.

The Department of Linguistic Science at Reading University was established in 1965 and was the first in Britain to offer a BA in Linguistics with postgraduate courses being offered in the following year. It is now one of the largest linguistics departments in Britain, with internationally renowned specialists in all major areas of the subject.

John Coleman reports on recent extensions to the nonconcatenative speech synthesis architecture employed in the YorkTalk and IPOX systems in *Synthesis of Connected Speech.*

Microsoft Speech API 4.0 SAPI Basics: This section introduces the eight main components of SAPI.

- Voice Command
- Voice Dictation
- Voice Text
- Voice Telephony
- Direct Speech Recognition
- Direct Text-to-Speech
- Audio Objects

The Speech Technology Group engages in research and development of spoken language technologies.

Some recent and notable publications from the Speech Technology group include the following:

> Whistler: A Trainable Text-to-Speech System, International Conference on Spoken Language Processing, 1996.
>
> If you would like to try the Whistler engine yourself, you can download it as part of the SAPI 4.0 Speech SDK.

DOWNLOAD: The Microsoft Speech SDK 4.0a

Also, make sure you understand how to buy a microphone for use with these downloads. I've also assembled some handy tips on problems with sound cards, microphones, and speakers.

Microsoft is currently working on a new release of the Microsoft Speech API V.5.0. This document addresses all the most frequently asked questions (FAQ).

What is the AT&T Advanced Speech Products Group?

The AT&T Advanced Speech Products Group offers software and hardware-based speech recognition and synthesis, speech coding, and audio coding technology platforms. These platforms can be integrated into many third-party applications and hardware configurations to provide speech-enabled products and services that meet the needs of a broad range of customers. The AT&T Advanced Speech Products Group primarily serves other organizations within the AT&T community.

- AT&T WATSON—for a general overview of AT&T Lab's Speech Technologies
- TTS Demo—on-line access to AT&T Lab's Next Generation Text-to-Speech

---

The Bell Labs text-to-speech system (TTS) has various applications including reading electronic mail messages, generating spoken prompts in voice response systems, and as an interface to an order-verification system for salespeople in the field.

They have a new book describing their work on multilingual text-to-speech: *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach.*

Lucent's text-to-speech engine (TTS) is the best text-to-speech currently available. The current engine and API are available from single- and multi-line packages for deployment on single PCs, all the way up to heavy-duty network embedded applications.

LTTS 3.1 is the cumulative product of many years of research by a large team of researchers under the direction of Bell Labs veteran, Joseph Olive, Ph.D., a physicist and composer. The system architecture in brief: input text is subjected to several phases of grammatical analysis, expansion of abbreviations, and heuristic normalizations (e.g., rules that determine how, for example, large numbers are properly read aloud) by a source-language-specific parser. The pre-processed output is used to index into a library of diphone samples, producing a basic

waveform table. Waveform data is then subjected to signal processing to impose simulated vocal-tract characteristics and appropriate prosody—the later determined by earlier grammatical analysis and optionally inflected by the programmer. More can be found here and at http://www.computertelephony.com/.

You can get the latest specifications for SABLE, an SGML-based TTS Markup language, or play around with a demo of a predecessor to SABLE, STML, here.

Also, look at the following link just for fun: English/Pig Latin "Translator"

After you get past playing with the web page, developer kits are available starting at $595 for a single-channel, host-based engine running under Windows on a 133-MHz Pentium. It is a case of "you get what you pay for" in this area of technology.

SounText is a high-quality low-cost multi-lingual speech synthesizer for MS-DOS and Microsoft Windows environments.

The standard package supports English, French, German, Italian and Spanish using Berkeley Speech Technology. Mandarin Chinese is also available.

Royal Society for the Blind of South Australia Adaptive Technology Centre web site.

American Foundation for the Blind.

Here is a tidbit of wisdom to chew on while your designing your high-power, slow, graphics site: *The Applicability of the Americans with Disabilities Act to the Internet.*

I've had more compliments from people with disabilities about my own site being "friendly" to them, while only a couple of people have ever told me that the site "looks like something done with Mosaic in the 80's."

VoiceXML Forum—Bringing voice access to the Web!

Launched in April of 2000 from the Lucent Technologies New Ventures Group, face2face, used years of Bell Labs research to create an innovative new software suite, which will revolutionize facial animation and lip synchronization for film and television production, electronic gaming, and the Internet.

The Perceptual Science Laboratory is engaged in a variety of experimental and theoretical inquiries in perception and cognition. A major research area concerns speech perception by ear, eye, and facial animation. They tested a general fuzzy logical model of perception in a variety of domains, including perception and understanding of language, memory, object, shape and depth perception, learning, and decision making. Research is also being carried out in reading.

Check their extensive lists of links to similar research.

ReadPlease 2000 shatters the myth that computers must sound robotic and monotonous. Just imagine having web pages and e-mail read aloud to you.

TimeTalk is a free demonstration of the customized female voice text-to-speech brought to you by Fonix Corporation.

TimeTalk is a clock utility that runs in your system tray and uses Fonix customized text-to-speech to announce the time.

Fonix claims TimeTalk is an example of the best sounding text-to-speech in the world today. I hope their web site is not representative of their quality; not a single link worked when I tried it. They claim to offer several other speech products, but I could not find anything about them.

HANDBOOK of Standards and Resources for Spoken Language Systems covers the EAGLES project, which is structured into five working groups on Text Corpora, Computational Lexicons, Computational Linguistic Formalisms, Evaluation, and Spoken Language.

Des Gestes Ecrits Aux Gestes Parles by A.I.C. Monaghan covers speech gestrures and their comparison with text gestures.

Abstract:

> All speech is gesture. Gestures of the tongue, lips and jaw make distinctions between different vowels and consonants. These are overlaid on gestures determining voice quality, pitch and loudness.
>
> Text can also be seen as a sequence of overlapping gestures. The use of underlining, **bolding**, *italics*, indentation, "quotation marks" and other annotations in rich text or hypertext formats corresponds to gestures in spoken communication.

Simtel.Net always has something fitting for my Resource Pages:

voicess.zip Veritel Voice Authentication Screen Saver. Free

Wave To Text v2.0 (a Voice Explorer series) is an English language speech-recognition-based dictation pad with a Wave To Text Wizard. The dictation pad converts in real time your voice to text, while the wizard converts a off-line recorded Windows Wave file containing continuous speech to English text. It writes the text to its own pad, and from there you can easily transfer it via a simple cut-and-paste operation to wherever you want.

Special requirements: Sound card, PC microphone, VB 5 runtimes (available from Simtel.Net as vb500a.zip).

Shareware.

Sandeep Thite, United Research Labs
unitedresearch@vsnl.com
http://www.research-lab.com/

I never like to do a Resource Page without throwing in something off the beaten path.

TruthVSA: Voice Stress Analysis Freeware

> "In both principle and execution VSA is a simple technology. Researchers found frequencies in the human voice in the 8 to 12 Hz range are sensitive to honesty. When a person is being honest the average sound in that range is generally below 10 Hz, but is usually above 10 Hz in dishonest situations." [I've seen these referred to as Microtremors.]
>
> "TruthVSA is a simple program which takes digital audio files as input, and outputs new ones with a changing tone in the backgrounds indicating the changing stress levels. Higher tones mean higher stress. It has one control: a threshold setting which determines how high the voice stress frequency must be to trigger the background tone. It also outputs a text log file giving a breakdown of the VSA data processed in each file. Programmers interested in developing more complex VSA applications will find complete [source code] included in the zip file."
>
> > "...This is where the art comes in; the operator has to learn to recognize patterns of stress and has to know something about the psychology of honest and dishonest people to read VSA results accurately. Although TVSA3 is freeware and anyone with a properly equipped computer can use it, it's not a tool for the inexperienced, judgmental or sloppy." - *Mike Kemp: Snitch Detector.*

I often think this would make an interesting *Circuit Cellar* project. It seems like it would be easy to do with todays DSP chips. If anyone does, let me know. I want one.

If your interested in this kind of thing, you might want to see what *The American Polygraph Association* has to say.

All product names and logos contained herein are the trademarks of their respective holders.

The fact that an item is listed here does not mean we promote its use for your application. No endorsement of the vendor or product is made or implied.

If you would like to add any information on this topic or request a specific topic to be covered, contact Bob Paddock.

Circuit Cellar provides up to date information for engineers, www.circuitcellar.com for more information and additional articles.

For subscription information, call (860) 875-2199 or e-mail subscribe@circuitcellar.com