Voice Technology

Nya Dover

First Edition, 2012

ISBN 978-81-323-4261-8

Published by:
White Word Publications
4735/22 Prakashdeep Bldg,
Ansari Road, Darya Ganj,
Delhi - 110002
Email: info@wtbooks.com

Table of Contents

Chapter 1 - Speech Coding and Electroglottograph

Chapter 2 - Psychoacoustics

Chapter 3 - Electronic Fluency Devices

Chapter 4 - Microsoft Speech API

Chapter 5 - Voice Analysis and Speaker Recognition

Chapter 6 - Voice Stress Analysis

Chapter 7 - Speech Recognition

Chapter 8 - Speech Synthesis

Chapter 9 - Voicemail

Chapter 10 - Voice Over IP

Chapter 1

Speech Coding and Electroglottograph

Speech coding

Speech coding is the application of data compression of digital audio signals containing speech. Speech coding uses speech-specific parameter estimation using audio signal processing techniques to model the speech signal, combined with generic data compression algorithms to represent the resulting modeled parameters in a compact bitstream.

The two most important applications of speech coding are mobile telephony and Voice over IP.

The techniques used in speech coding are similar to that in audio data compression and audio coding where knowledge in psychoacoustics is used to transmit only data that is relevant to the human auditory system. For example, in voiceband speech coding, only information in the frequency band 400 Hz to 3500 Hz is transmitted but the reconstructed signal is still adequate for intelligibility.

Speech coding differs from other forms of audio coding in that speech is a much simpler signal than most other audio signals, and much a lot more statistical information is available about the properties of speech. As a result, some auditory information which is relevant in audio coding can be unnecessary in the speech coding context. In speech coding, the most important criterion is preservation of intelligibility and "pleasantness" of speech, with a constrained amount of transmitted data.

It should be emphasised that the intelligibility of speech includes, besides the actual literal content, also speaker identity, emotions, intonation, timbre etc. that are all important for perfect intelligibility. The more abstract concept of pleasantness of degraded speech is a different property than intelligibility, since it is possible that degraded speech is completely intelligible, but subjectively annoying to the listener.

In addition, most speech applications require low coding delay, as long coding delays interfere with speech interaction.

Sample companding viewed as a form of speech coding

From this viewpoint, the A-law and μ-law algorithms (G.711) used in traditional PCM digital telephony can be seen as a very early precursor of speech encoding, requiring only 8 bits per sample but giving effectively 12 bits of resolution. Although this would generate unacceptable distortion in a music signal, the peaky nature of speech waveforms, combined with the simple frequency structure of speech as a periodic waveform with a single fundamental frequency with occasional added noise bursts, make these very simple instantaneous compression algorithms acceptable for speech.

A wide variety of other algorithms were tried at the time, mostly variants on delta modulation, but after careful consideration, the A-law/μ-law algorithms were chosen by the designers of the early digital telephony systems. At the time of their design, their 33% bandwidth reduction for a very low complexity made them an excellent engineering compromise. Their audio performance remains acceptable, and there has been no need to replace them in the stationary phone network.

In 2008, G.711.1 codec, which has a scalable structure, was standardized by ITU-T. The input sampling rate is 16 kHz.

Modern speech compression

Much of the later work in speech compression was motivated by military research into digital communications for secure military radios, where very low data rates were required to allow effective operation in a hostile radio environment. At the same time, far more processing power was available, in the form of VLSI integrated circuits, than was available for earlier compression techniques. As a result, modern speech compression algorithms could use far more complex techniques than were available in the 1960s to achieve far higher compression ratios.

These techniques were available through the open research literature to be used for civilian applications, allowing the creation of digital mobile phone networks with substantially higher channel capacities than the analog systems that preceded them.

The most common speech coding scheme is Code Excited Linear Prediction (CELP) coding, which is used for example in the GSM standard. In CELP, the modelling is divided in two stages, a linear predictive stage that models the spectral envelope and code-book based model of the residual of the linear predictive model.

In addition to the actual speech coding of the signal, it is often necessary to use channel coding for transmission, to avoid losses due to transmission errors. Usually, speech coding and channel coding methods have to be chosen in pairs, with the more important

bits in the speech data stream protected by more robust channel coding, in order to get the best overall coding results.

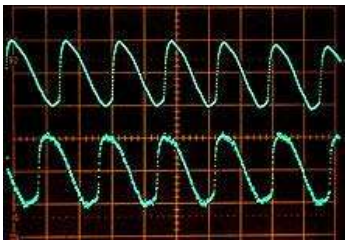
The Speex project is an attempt to create a free software speech coder, unencumbered by patent restrictions.

Major subfields:

- Wide-band speech coding
  - AMR-WB for WCDMA networks
  - VMR-WB for CDMA2000 networks
  - G.722, G.722.1, Speex and others for VoIP and videoconferencing
- Narrow-band speech coding
  - FNBDT for military applications
  - SMV for CDMA networks
  - Full Rate, Half Rate, EFR, AMR for GSM networks
  - G.723.1, G.726, G.728, G.729, iLBC and others for VoIP or videoconferencing

Electroglottograph

Electroglottograph, Glottal Enterprises model EG2-PCX shown here.



Photograph of an EGG signal from a Glottal Enterprises EG2-PC (top) and a Laryngograph/Kay electroglottograph (bottom).

Showing the contacts on the electrodes from a Glottal Enterprises EG2-PCX. Electrodes for other electroglottographs are typically very similar in size and shape. This set of electrodes is from a Glottal Enterprises EG2-PCX, which is a dual-channel EGG, so it has 2 sets of contacts. Electrode jelly is used to help conduct the signal from the contacts to the neck.

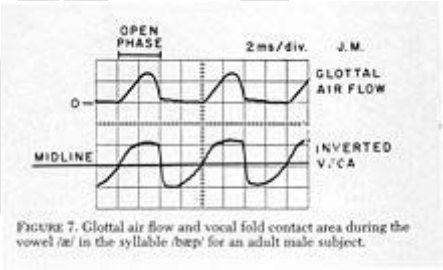


FIGURE 7. Glottal air flow and vocal fold contact area during the vowel /æ/ in the syllable /bæp/ for an adult male subject.

Figure 7 of: Martin Rothenberg and James J. Mashie, *Monitoring Vocal Fold Abduction Through Vocal Fold Contact Area* Journal of Speech and Hearing Research, Volume 31, 338-351, September 1988

The **electroglottograph**, or **EGG**, (sometimes referred to as a laryngograph) is a device for the noninvasive measurement of the time variation of the degree of contact between the vibrating vocal folds during voice production. Though it is difficult to verify the assumption precisely, the aspect of contact being measured by a typical EGG unit is considered to be the vocal fold contact area (VFCA). To measure VFCA, an EGG records variations in the transverse electrical impedance of the larynx and nearby tissues by means of a small A/C electrical current in the megaHertz region applied by electrodes on the surface of the neck. This electrical impedance will vary slightly with the area of contact between the moist vocal folds during that part of the glottal vibratory cycle in which the folds are in contact. However, because the percentage variation in the neck impedance caused by vocal fold contact can be extremely small and varies considerably between subjects, no absolute measure of contact area is obtained, only the pattern of variation for a given subject.

Early commercial available EGG units were compared quite thoroughly by Baken. However, using modern low noise electronics, EGG noise levels can be brought down enough so that the noise is approximately 40dB (a factor of 100) less than a typical EGG signal from an adult voice.

In addition, by the use of multiple channels simultaneously, the technique can be made easier to use and more reliable by giving the user an indication of the correct positioning of the electrodes, and providing a quantitative measure of vertical movements of the larynx during voice production.

Electroglottograph signals have found use in stroboscope synchronization, voice fundamental frequency tracking, tracking vocal fold abductory movements and the study of the singing voice.

Chapter 2

Psychoacoustics

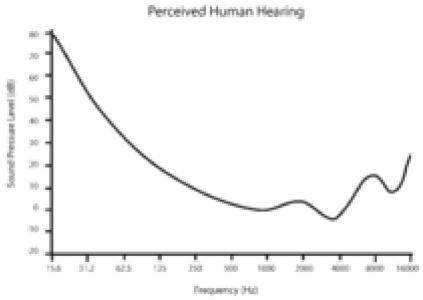
Psychoacoustics is the scientific study of sound perception. More specifically, it is the branch of science studying the psychological and physiological responses associated with sound (including speech and music). It can be further categorized as a branch of Psychophysics.

Background

Hearing is not a purely mechanical phenomenon of wave propagation, but is also a sensory and perceptual event; in other words, when a person hears something, that something arrives at the ear as a mechanical sound wave traveling through the air, but within the ear it is transformed into neural action potentials. These nerve pulses then travel to the brain where they are perceived. Hence, in many problems in acoustics, such as for audio processing, it is advantageous to take into account not just the mechanics of the environment, but also the fact that both the ear and the brain are involved in a person's listening experience.

The inner ear, for example, does significant signal processing in converting sound waveforms into neural stimulus, so certain differences between waveforms may be imperceptible. Audio compression techniques, such as MP3, make use of this fact. In addition, the ear has a nonlinear response to sounds of different loudness levels. Telephone networks and audio noise reduction systems make use of this fact by nonlinearly compressing data samples before transmission, and then expanding them for playback. Another effect of the ear's nonlinear response is that sounds that are close in frequency produce phantom beat notes, or intermodulation distortion products.

Limits of perception



An equal-loudness contour. Note peak sensitivity between 2khz and 4khz, the frequency around which the human voice centers

The human ear can nominally hear sounds in the range 20 Hz to 20,000 Hz (20 kHz). This upper limit tends to decrease with age, most adults being unable to hear above 16 kHz. The ear itself does not respond to frequencies below 20 Hz, but these can be perceived via the body's sense of touch.

Frequency resolution of the ear is 3.6 Hz within the octave of 1,000–2,000 Hz. That is, changes in pitch larger than 3.6 Hz can be perceived in a clinical setting. However, even smaller pitch differences can be perceived through other means. For example, the interference of two pitches can often be heard as a (low-)frequency difference pitch. This effect of phase variance upon the resultant sound is known as beating.

The semitone scale used in Western musical notation is not a linear frequency scale but logarithmic. Other scales have been derived directly from experiments on human hearing perception, such as the mel scale and Bark scale (these are used in studying perception, but not usually in musical composition), and these are approximately logarithmic in frequency at the high-frequency end, but nearly linear at the low-frequency end.

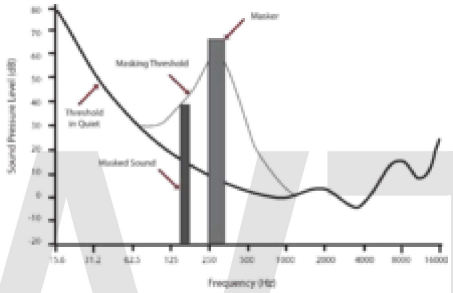
The intensity range of audible sounds is enormous. Our ear drums are sensitive only to variations in the sound pressure, but can detect pressure changes as small as 2×10^{-10} atm and as great as or greater than 1 atm. For this reason, sound pressure level is also measured logarithmically, with all pressures referenced to 1.97385×10^{-10} atm. The lower limit of audibility is therefore defined as 0 dB, but the upper limit is not as clearly defined. The upper limit is more a question of the limit where the ear will be physically harmed or with the potential to cause noise-induced hearing loss.

A more rigorous exploration of the lower limits of audibility determines that the minimum threshold at which a sound can be heard is frequency dependent. By measuring this minimum intensity for testing tones of various frequencies, a frequency dependent absolute threshold of hearing (ATH) curve may be derived. Typically, the ear shows a peak of sensitivity (i.e., its lowest ATH) between 1 kHz and 5 kHz, though the threshold changes with age, with older ears showing decreased sensitivity above 2 kHz.

The ATH is the lowest of the equal-loudness contours. Equal-loudness contours indicate the sound pressure level (dB), over the range of audible frequencies, which are perceived as being of equal loudness. Equal-loudness contours were first measured by Fletcher and Munson at Bell Labs in 1933 using pure tones reproduced via headphones, and the data they collected are called Fletcher-Munson curves. Because subjective loudness was difficult to measure, the Fletcher-Munson curves were averaged over many subjects.

Robinson and Dadson refined the process in 1956 to obtain a new set of equal-loudness curves for a frontal sound source measured in an anechoic chamber. The Robinson-Dadson curves were standardized as ISO 226 in 1986. In 2003, ISO 226 was revised as equal-loudness contour using data collected from 12 international studies.

Masking effects



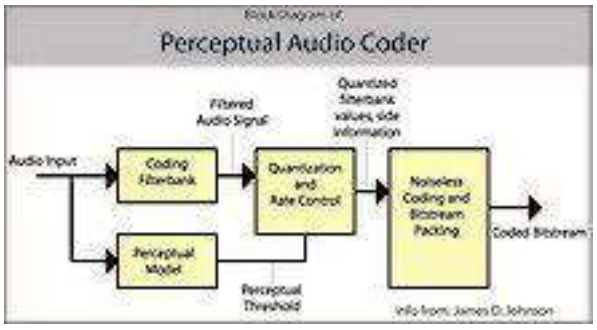
Audio Masking Graph

In some situations an otherwise clearly audible sound can be masked by another sound. For example, conversation at a bus stop can be completely impossible if a loud bus is driving past. This phenomenon is called masking. A weaker sound is masked if it is made inaudible in the presence of a louder sound.

Missing fundamental

A harmonic series of pitches that are related 2×f, 3×f, 4×f, 5×f, etc., give human hearing the psychoacoustic impression that the pitch 1×f is present.

Software



Perceptual Audio Coding uses the Psychoacoustics algorithm

The **psychoacoustic model** provides for high quality lossy signal compression by describing which parts of a given digital audio signal can be removed (or aggressively compressed) safely — that is, without significant losses in the (consciously) perceived quality of the sound.

It can explain how a sharp clap of the hands might seem painfully loud in a quiet library, but is hardly noticeable after a car backfires on a busy, urban street. This provides great benefit to the overall compression ratio, and psychoacoustic analysis routinely leads to compressed music files that are 1/10 to 1/12 the size of high quality masters with very little discernible loss in quality. Such compression is a feature of nearly all modern audio compression formats. Some of these formats include Dolby Digital (AC-3), MP3, Ogg Vorbis, AAC, WMA, MPEG-1 Layer II (used for digital audio broadcasting in several countries) and ATRAC, the compression used in MiniDisc and some Walkman models.

Psychoacoustics is based heavily on human anatomy, especially the ear's limitations in perceiving sound as outlined previously. To summarize, these limitations are:

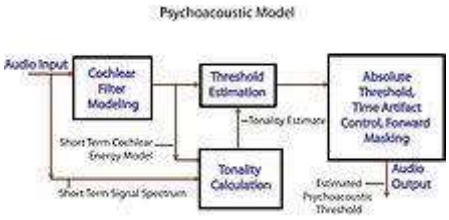
- High frequency limit
- Absolute threshold of hearing
- Temporal masking
- Simultaneous masking

Given that the ear will not be at peak perceptive capacity when dealing with these limitations, a compression algorithm can assign a lower priority to sounds outside the range of human hearing. By carefully shifting bits away from the unimportant components and toward the important ones, the algorithm ensures that the sounds a listener is most likely to perceive are of the highest quality.

Music

Psychoacoustics include topics and studies which are relevant to music psychology and music therapy. Theorists such as Benjamin Boretz consider some of the results of psychoacoustics to be meaningful only in a musical context.

Applied psychoacoustics



Psychoacoustics Model

Psychoacoustics is presently applied within many fields from software development, where developers map proven and experimental mathematical patterns; in digital signal

processing, where many audio compression codecs such as MP3 use a psychoacoustic model to increase compression ratios; in the design of (high end) audio systems for accurate reproduction of music in theatres and homes; as well as defense systems where scientists have experimented with limited success in creating new acoustic weapons, which emit frequencies that may impair, harm, or kill. It is also applied today within music, where musicians and artists continue to create new auditory experiences by masking unwanted frequencies of instruments, causing other frequencies to be enhanced. Yet another application is in design of small or lower-quality loudspeakers, which use the phenomenon of missing fundamentals to give the effect of low frequency bass notes that the system, due to frequency limitations, cannot actually reproduce.

Chapter 3

Electronic Fluency Devices

Electronic fluency devices (also known as **assistive devices**, **electronic aids**, **altered auditory feedback devices** and **altered feedback devices**) are electronic devices intended to improve the fluency of persons who stutter. Most electronic fluency devices change the sound of the user's voice in his or her ear.

Electronic fluency device

Types

Electronic fluency devices can be divided into two basic categories.

- Computerized feedback devices provide feedback on the physiological control of respiration and phonation, including loudness, vocal intensity and breathing patterns.

- Altered auditory feedback (AAF) devices alter the speech signal so that speakers hear their voices differently.

Computerized feedback devices

Computerized feedback devices (such as CAFET or Dr. Fluency) use computer technology to increase control over breathing and phonation. A microphone gathers information about the stutterer's speech and feedback is delivered on a computer screen.

Measurements include intensity (loudness), voice quality, breathing patterns, and voicing strategies. These programs are designed to train features related to prolonged speech, a treatment technique which is frequently used in stuttering therapy. No peer-reviewed studies have been published showing the effectiveness of commercial systems in a clinical context. A study of electromyographic (EMG) feedback in children and adolescents found it to be as effective as other treatments (home-based and clinic-based smooth speech training) in the short and longterm.

Altered auditory feedback devices

Altered auditory feedback (AAF) such as singing, choral speaking, masking, delayed or frequency altered feedback have long been known to reduce stuttering. Early altered auditory feedback devices were large and thus confined to the laboratory or therapy room, but advances in electronics have permitted increasingly portable devices such as Derazne Correctophone, the Edinburgh Masker, the Vocaltech Clinical Vocal Feedback Device, the Fluency Master and the SpeechEasy. Current devices may be similar in size and appearance to a hearing aid, including in-the-ear and completely-in-the-canal models.

Masking

White noise masking has been well documented to reduce stuttering. Clinic-based and portable devices, such as the Edinburgh Masker (since discontinued) have been developed to deliver masking, and found that masking was effective in reducing stuttering, though many found that reduction in stuttering faded with time. Interest in masking reduced during the 1980s as a result of studies finding delayed auditory feedback and frequency altered feedback were more effective in reducing stuttering.

Delayed auditory feedback

The effect of delayed auditory feedback (DAF) in reducing stuttering has been noted since the 1950s. A DAF user hears his or her voice in headphones, delayed a fraction of a second. Typical delays are in the 50 millisecond to 200 millisecond range. In stutterers, DAF may produce slow, prolonged but fluent speech. In the 1960s to 1980s, DAF was mainly used to train prolongation and fluency. As the stutterer masters fluent speech skills at a slow speaking rate, the delay is reduced in stages, gradually increasing speaking rate, until the person can speak fluently at a normal speaking rate. It was not until the 1990s that research began to focus on DAF in isolation. Recent studies have moved from longer delays to shorter delays in the 50 millisecond to 75 millisecond range, and have found that speakers can maintain fast rates and achieve increased fluency at these delays. Delayed auditory feedback presented binaurally (i.e. in both ears) is more effective than that presented in monaurally, or in one ear only.

Frequency-altered feedback

Pitch-shifting frequency-altered auditory feedback (FAF) changes the pitch at which the user hears his or her voice. Varying pitch from quarter, half or full octave shift typically

results in 55–74% decreases stuttering in short reading tasks. Individuals differ as to direction and extent of the pitch shift required to maximally reduce stuttering. In studies that gave longer exposure to FAF and used more meaningful daily life tasks such as generating a monologue, only some participants experienced a reduction in stuttering. Initial claims that AAF was more powerful than FAF in reducing stuttering have not been supported by subsequent research. FAF is, like DAF, more effective when presented binaurally.

Effectiveness

Studies have shown that altered auditory feedback (including delayed auditory feedback, frequency altered feedback) as provided by devices such as the Casa Futura School DAF machine or SpeechEasy can immediately reduce stuttering by 40 to 80 per cent in reading tasks. Laboratory studies suggest that reductions in stuttering with an electronic fluency device can occur without a reduced speech rate, and that speech naturalness is often enhanced with AAF. However, the effects of altered feedback are highly individualistic, with some obtaining considerable increases in fluency, while others receive little or no benefit.

A 2006 review of stuttering treatments noted that none of the treatment studies on altered auditory feedback met the criteria for experimental quality. In addition, studies have been critiqued for failing to demonstrate ecological validity; in particular that AAF effects continue over the long term and in everyday speaking situations. The high-profile promotion in the media of devices such as the "SpeechEasy" has been criticized as inappropriate given the lack of scientific evidence for their effectiveness.

There are few published studies on the effect of the AAF in the daily activities of life; studies have mainly examined the effect of AAF on short oral reading tasks, with some studying the giving of a monologue that is usually short in duration. Several studies have produced group results that stutterers using the SpeechEasy show greater reductions in reading than for monologue and conversation. Using AAF was effective in reducing stuttering in scripted telephone calls and giving presentations according to two studies. Another study examining the effects of the SpeechEasy in more naturalistic situations (conversation and asking questions of strangers outside the clinic) found that the SpeechEasy failed to show a significant effect following 6 months of use, though individual subjects varied in their response. A further study examining the use of the device during phone and face to face conversation also found wide variations in stuttering reduction, with just under half exhibiting stable improvement over the course of the 4 months of the study.

While there is evidence of the immediate, short-term effectiveness of AAF devices in reducing stuttering, the longterm effects of altered feedback are unclear. There is some limited experimental data that in some speakers the effect of AAF may fade after a few minutes of exposure, and some anecdotal reports suggest that over time users receive continued but lessened effects from their device. While one group study has reported continued overall reductions in stuttering after a year of daily use of the SpeechEasy on

reading and a monologue task, others have found that some participants showed adaptation effects, gaining less benefit from the device after exposure for several months, including stuttering more with the device than without it. Some studies of various altered auditory feedback devices have noted carryover fluency, i.e. a reduction in stuttering after the stutterer removes an electronic fluency device, while others have not.

The effective of electronic fluency devices as measured by qualitative measures and ratings by stutterers have also been made. Studies show that some stutterers report improved fluency and confidence about speaking, and less severe stuttering and some carryover effects; the device is perceived as being particularly useful on the telephone. They reported that the device was difficult to use in noisy situations as the device amplifies all voices and sounds, and some acclimatization to the use of the device over time. Qualitative reports of satisfaction may be disassociated from more objective measures of fluency: some stutterers who gain little or no benefit from a device based on objective measures rate the device highly, while others who were obtaining benefit on measures of fluency reported negatives opinions about the device.

Use with children

There is little experimental evaluation of the therapeutic effect of AAF on children who stutter: one study noted that effects of FAF were less in children than adults. Given the lack of evidence of its effectiveness, as well as concerns about the impact of altered feedback on developing speech and language systems, some authors have expressed the view that the use of an AAF with children would be unethical.

Causes of altered auditory feedback effects

The precise reasons for the fluency-inducing effects of AAF in stutterers are unknown. Early investigators suggested that those who stutter had an abnormal speech–auditory feedback loop that was corrected or bypassed while speaking under DAF. Later researchers proposed increased fluency was actually caused by the changes in speech production, including slower speech rates, higher pitches and increased loudness, rather than the AAF per se. However, subsequent studies have noted that increased fluency occurred in some stutterers at normal and fast rates using DAF. Some suggest that stuttering is caused by defective auditory processing, and that AAF helps to correct the misperceived rhythmic structure of speech. It has been shown that some stutterers have noted that have atypical auditory anatomy and that DAF improved fluency in these stutterers but not in those with typical anatomy. However, positron emission tomography studies on choral reading in stutterers suggest that AAF also made changes in motor and speech production areas of the brain, as well as the auditory processing areas. Choral reading reduced the overactivity in motor areas that is found with stuttered reading, and largely reversed the left-hemisphere based auditory-system and speech production system underactivation. Noting that the effects of altered feedback vary from person to person and can wear off over time, distraction has also been proposed as a possible cause of stuttering reduction with AAF.

Chapter 4

Microsoft Speech API

The **Speech Application Programming Interface** or **SAPI** is an API developed by Microsoft to allow the use of speech recognition and speech synthesis within Windows applications. To date, a number of versions of the API have been released, which have shipped either as part of a Speech SDK, or as part of the Windows OS itself. Applications that use SAPI include Microsoft Office, Microsoft Agent and Microsoft Speech Server.

In general all versions of the API have been designed such that a software developer can write an application to perform speech recognition and synthesis by using a standard set of interfaces, accessible from a variety of programming languages. In addition, it is possible for a 3rd-party company to produce their own Speech Recognition and Text-To-Speech engines or adapt existing engines to work with SAPI. In principle, as long as these engines conform to the defined interfaces they can be used instead of the Microsoft-supplied engines.

In general the Speech API is a freely-redistributable component which can be shipped with any Windows application that wishes to use speech technology. Many versions (although not all) of the speech recognition and synthesis engines are also freely redistributable.

There have been two main 'families' of the Microsoft Speech API. SAPI versions 1 through 4 are all similar to each other, with extra features in each newer version. SAPI 5 however was a completely new interface, released in 2000. Since then several sub-versions of this API have been released.

Basic architecture

Broadly the Speech API can be viewed as an interface or piece of middleware which sits between *applications* and speech *engines* (recognition and synthesis). In SAPI versions 1 to 4, applications could directly communicate with engines. The API included an abstract *interface definition* which applications and engines conformed to. Applications could also use simplified higher-level objects rather than directly call methods on the engines.

In SAPI 5 however, applications and engines do not directly communicate with each other. Instead each talk to a runtime component (**sapi.dll**). There is an API implemented by this component which applications use, and another set of interfaces for engines.

Typically in SAPI 5 applications issue calls through the API (for example to load a recognition grammar; start recognition; or provide text to be synthesized). The sapi.dll runtime component interprets these commands and processes them, where necessary calling on the engine through the engine interfaces (for example, the loading of a grammar from a file is done in the runtime, but then the grammar data is passed to the recognition engine to actually use in recognition). The recognition and synthesis engines also generate events while processing (for example, to indicate an utterance has been recognized or to indicate word boundaries in the synthesized speech). These pass in the reverse direction, from the engines, through the runtime dll, and on to an *event sink* in the application.

In addition to the actual API definition and runtime dll, other components are shipped with all versions of SAPI to make a complete Speech Software Development Kit. The following components are among those included in most versions of the Speech SDK:

- *API definition files* - in MIDL and as C or C++ header files.
- *Runtime components* - e.g. sapi.dll.
- *Control Panel applet* - to select and configure default speech recognizer and synthesizer.
- *Text-To-Speech engines* in multiple languages.
- *Speech Recognition engines* in multiple languages.
- *Redistributable components* to allow developers to package the engines and runtime with their application code to produce a single installable application.
- *Sample application code*.
- *Sample engines* - implementations of the necessary engine interfaces but with no true speech processing which could be used as a sample for those porting an engine to SAPI.
- *Documentation*.

Versions

Xuedong Huang was a key person who led Microsoft's early SAPI efforts.

SAPI 1-4 API family

SAPI 1

The first version of SAPI was released in 1995, and was supported on Windows 95 and Windows NT 3.51. This version included low-level Direct Speech Recognition and Direct Text To Speech APIs which applications could use to directly control engines, as well as simplified 'higher-level' Voice Command and Voice Talk APIs.

SAPI 2

SAPI 2.0 was released in 1996.

SAPI 3

SAPI 3.0 was released in 1997. It added limited support for dictation speech recognition (discrete speech, not continuous), and additional sample applications and audio sources.

SAPI 4

SAPI 4.0 was released in 1998. This version of SAPI included both the core COM API; together with C++ wrapper classes to make programming from C++ easier; and ActiveX controls to allow drag-and-drop Visual Basic development. This was shipped as part of an SDK that included recognition and synthesis engines. It also shipped (with synthesis engines only) in Windows 2000.

The main components of the SAPI 4 API (which were all available in C++, COM, and ActiveX flavors) were:

- **Voice Command** - high-level objects for command & control speech recognition
- **Voice Dictation** - high-level objects for continuous dictation speech recognition
- **Voice Talk** - high-level objects for speech synthesis
- **Voice Telephony** - objects for writing telephone speech applications
- **Direct Speech Recognition** - objects for direct control of recognition engine
- **Direct Text To Speech** - objects for direct control of synthesis engine
- **Audio objects** - for reading to and from an audio device or file

SAPI 5 API family

The **Speech SDK version 5.0**, incorporating the **SAPI 5.0** runtime was released in 2000. This was a complete redesign from previous versions and neither engines nor applications which used older versions of SAPI could use the new version without considerable modification.

The design of the new API included the concept of strictly separating the application and engine so all calls were routed through the runtime sapi.dll. This change was intended to make the API more 'engine-independent', preventing applications from inadvertently depending on features of a specific engine. In addition this change was aimed at making it much easier to incorporate speech technology into an application by moving some management and initialization code into the runtime.

The new API was initially a pure COM API and could be used easily only from C/C++. Support for VB and scripting languages were added later. Operating systems from Windows 98 and NT 4.0 upwards were supported.

Major features of the API include:

- **Shared Recognizer**. For desktop speech recognition applications, a recognizer object can be used that runs in a separate process (**sapisvr.exe**). All applications using the shared recognizer communicate with this single instance. This allows sharing of resources, removes contention for the microphone and allows for a global UI for control of all speech applications.
- **In-proc recognizer**. For applications that require explicit control of the recognition process the in-proc recognizer object can be used instead of the shared one.
- **Grammar objects**. Speech grammars are used to specify the words that the recognizer is listening for. SAPI 5 defines an XML markup for specifying a grammar, as well as mechanisms to create them dynamically in code. Methods also exist for instructing the recognizer to load a built-in dictation language model.
- **Voice object**. This performs speech synthesis, producing an audio stream from text. A markup language (similar to XML, but not strictly XML) can be used for controlling the synthesis process.
- **Audio interfaces**. The runtime includes objects for performing speech input from the microphone or speech output to speakers (or any sound device); as well as to and from wave files. It is also possible to write a custom audio object to stream audio to or from a non-standard location.
- **User lexicon object**. This allows custom words and pronunciations to be added by a user or application. These are added to the recognition or synthesis engine's built-in lexicons.
- **Object tokens**. This is a concept allowing recognition and TTS engines, audio objects, lexicons and other categories of object to be registered, enumerated and instantiated in a common way.

SAPI 5.0

This version shipped in late 2000 as part of the Speech SDK version 5.0, together with version 5.0 recognition and synthesis engines. The recognition engines supported continuous dictation and command & control and were released in U.S. English, Japanese and Simplified Chinese versions. In the U.S. English system, special acoustic models were available for children's speech and telephony speech. The synthesis engine was available in English and Chinese. This version of the API and recognition engines also shipped in Microsoft Office XP in 2001.

SAPI 5.1

This version shipped in late 2001 as part of the Speech SDK version 5.1. Automation-compliant interfaces were added to the API to allow use from Visual Basic, scripting languages such as JScript, and managed code. This version of the API and TTS engines was shipped in Windows XP. Windows XP Tablet PC Edition and Office 2003 also

include this version, but with a substantially improved version 6 recognition engine and Traditional Chinese.

SAPI 5.2

This was a special version of the API for use only in the Microsoft Speech Server which shipped in 2004. It added support for SRGS and SSML mark-up languages, as well as additional server features and performance improvements. The Speech Server also shipped with the version 6 desktop recognition engine and the version 7 server recognition engine.

SAPI 5.3

This is the version of the API that ships in Windows Vista together with new recognition and synthesis engines. As Windows Speech Recognition is now integrated into the operating system, the Speech SDK and APIs are a part of the Windows SDK. SAPI 5.3 includes the following new features:

- Support for W3C XML speech grammars for recognition and synthesis. The Speech Synthesis Markup Language (SSML) version 1.0 provides the ability to mark up voice characteristics, speed, volume, pitch, emphasis, and pronunciation.
- The Speech Recognition Grammar Specification (SRGS) supports the definition of context-free grammars, with two limitations:
    - It does not support the use of SRGS to specify dual-tone modulated-frequency (touch-tone) grammars.
    - It does not support Augmented Backus–Naur form (ABNF).
- Support for semantic interpretation script within grammars. SAPI 5.3 enables an SRGS grammar to be annotated with JavaScript for semantic interpretation to supplement the recognized text.
- User-Specified shortcuts in lexicons, which is the ability to add a string to the lexicon and associate it with a shortcut word. When dictating, the user can say the shortcut word and the recognizer will return the expanded string.
- Additional functionality and ease-of-programming provided by new types.
- Performance improvements, improved reliability and security.
- Version 8 of the speech recognition engine ("Microsoft Speech Recognizer")

SAPI 5.4

This is an updated version of the API that ships in Windows 7.

SAPI 5 Voices

Microsoft Sam (Speech Articulation Module) is a commonly-shipped SAPI 5 voice. In addition, Microsoft Office XP and Office 2003 installed L&H Michael and Michelle voices. The SAPI 5.1 SDK installs 2 more voices, *Mike* and *Mary*. Windows Vista includes Microsoft Anna which replaces Microsoft Sam. Anna is designed to sound more

natural and offer greater intelligibility. The Chinese version of Windows Vista and later Windows client versions also include a female voice named Microsoft Lili. Microsoft Anna is also installed on Windows XP by Microsoft Streets & Trips 2006 and later versions.

Managed code Speech API

A managed code API ships as part of the .NET Framework 3.0. It has similar functionality to SAPI 5 but is more suitable to be used by managed code applications. The new API is available on Windows XP, Windows Server 2003, Windows Vista, and Windows Server 2008.

The existing SAPI 5 API can also be used from managed code to a limited extent by creating COM Interop code (helper code designed to assist in accessing COM interfaces and classes). This works well in some scenarios however the new API should provide a more seamless experience equivalent to using any other managed code library.

Speech functionality in Windows Vista

Windows Vista includes a number of new speech-related features including:

- Speech control of the full Windows GUI and applications
- New tutorial, microphone wizard, and UI for controlling speech recognition
- New version of the Speech API runtime: SAPI 5.3
- Built-in updated Speech Recognition engine (Version 8)
- New Speech Synthesis engine and SAPI voice Microsoft Anna
- Managed code speech API (codenamed SpeechFX)
- Speech recognition support for 8 languages at release time: U.S. English, U.K. English, traditional Chinese, simplified Chinese, Japanese, German, French and Spanish, with more language to be released later.
- Microsoft Agent most notably, and all other Microsoft speech applications use SAPI 5.

Compatibility

The Speech API is compatible with the following operating systems:

SAPI 5

- Microsoft Windows 7
- Microsoft Windows Vista
- Microsoft Windows 2003
- Microsoft Windows XP
- Microsoft Windows 2000

SAPI 4

- Microsoft Windows Millennium Edition
- Microsoft Windows 98
- Microsoft Windows NT 4.0, Service Pack 6a, in English, Japanese and Simplified Chinese.

Major applications using SAPI

- Microsoft Windows XP Tablet PC Edition includes SAPI 5.1 and speech recognition engines 6.1 for English, Japanese, and Chinese (simplified and traditional)
- Windows Speech Recognition in Windows Vista
- Microsoft Narrator in Windows 2000 and later Windows operating systems
- Microsoft Office XP and Office 2003
- Microsoft Excel 2002, Microsoft Excel 2003, and Microsoft Excel 2007 for speaking spreadsheet data
- Microsoft Voice Command for Windows Pocket PC and Windows Mobile
- Microsoft Plus! Voice Command for Windows Media Player
- Dragon NaturallySpeaking general-purpose speech recognition application
- Adobe Reader uses voice output to read document content
- CoolSpeech, text-to-speech application that reads text aloud from a variety of sources
- Window-Eyes screen reader
- JAWS screen reader
- NVDA open-source screen reader

Libraries using SAPI output

- FastFormat, via its speech_sink
- Pantheios, via its be.speech back-end

Chapter 5

Voice Analysis and Speaker Recognition

Voice analysis

Voice analysis is the study of speech sounds for purposes other than linguistic content, such as in speech recognition. Such studies include mostly medical analysis of the voice i.e. phoniatrics, but also speaker identification. More controversially, some believe that the truthfulness or emotional state of speakers can be determined using Voice Stress Analysis or Layered Voice Analysis.

Typical voice problems

A medical study of the voice can be, for instance, analysis of the voice of patients who have had a polyp removed from his or her vocal cords through an operation. In order to objectively evaluate the improvement in voice quality there has to be some measure of voice quality. An experienced voice therapist can quite reliably evaluate the voice, but this requires extensive training and is still always subjective.

Another active research topic in medical voice analysis is vocal loading evaluation. The vocal cords of a person speaking for an extended period of time will suffer from tiring, that is, the process of speaking exerts a load on the vocal cords where the tissue will suffer from tiring. Among professional voice users (i.e. teachers, sales people) this tiring can cause voice failures and sick leaves. To evaluate these problems vocal loading needs to be objectively measured.

Analysis methods

Voice problems that require voice analysis most commonly originate from the vocal folds or the laryngeal musculature that controls them, since the folds are subject to collision forces with each vibratory cycle and to drying from the air being forced through the small gap between them, and the laryngeal musclature is intensely active during speech or singing and is subject to tiring. However, dynamic analysis of the vocal folds and their movement is physically difficult. The location of the vocal folds effectively prohibits

direct, invasive measurement of movement. Less invasive imaging methods such as x-rays or ultrasounds do not work because the vocal cords are surrounded by cartilage which distort image quality. Movements in the vocal cords are rapid, fundamental frequencies are usually between 80 and 300 Hz, thus preventing usage of ordinary video. Stroboscopic, and high-speed videos provide an option but in order to see the vocal folds, a fiberoptic probe leading to the camera has to be positioned in the throat, which makes speaking difficult. In addition, placing objects in the pharynx usually triggers a gag reflex that stops voicing and closes the larynx. In addition, stroboscopic imaging is only useful when the vocal fold vibratory pattern is closely periodic.

The most important indirect methods are currently inverse filtering of either microphone or oral airflow recordings and electroglottography (EGG). In inverse filtering, the speech sound (the radiated acoustic pressure waveform, as obtained from a microphone) or the oral airflow waveform from a circumferentially vented (CV) mask is recorded outside the mouth and then filtered by a mathematical method to remove the effects of the vocal tract. This method produces an estimate of the waveform of the glottal airflow pulses, which in turn reflect the movements of the vocal folds. The other kind of noninvasive indirect indication of vocal fold motion is the electroglottography, in which electrodes placed on either side of the subject's throat at the level of the vocal folds record the changes in the conductivity of the throat according to how large a portion of the vocal folds are touching each other. It thus yields one-dimensional information of the contact area. Neither inverse filtering nor EGG are sufficient to completely describe the complex 3-dimensional pattern of vocal fold movement, but can provide useful indirect evidence of that movement.

Speaker recognition

Speaker recognition is the computing task of validating a user's claimed identity using characteristics extracted from their voices.

There is a difference between *speaker recognition* (recognizing **who** is speaking) and *speech recognition* (recognizing **what** is being said). These two terms are frequently confused, as is *voice recognition*. Voice recognition is combination of the two where it uses learned aspects of a speakers voice to determine what is being said - such a system cannot recognise speech from random speakers very accurately, but it can reach high accuracy for individual voices with which it has been trained. In addition, there is a difference between the act of authentication (commonly referred to as **speaker verification** or **speaker authentication**) and identification. Finally, there is a difference between *speaker recognition* (recognizing **who** is speaking) and *speaker diarisation* (recognizing **when** the **same** speaker is speaking).

Speaker recognition has a history dating back some four decades and uses the acoustic features of speech that have been found to differ between individuals. These acoustic patterns reflect both anatomy (e.g., size and shape of the throat and mouth) and learned behavioral patterns (e.g., voice pitch, speaking style). Speaker verification has earned speaker recognition its classification as a "behavioral biometric."

Verification versus identification

There are two major applications of *speaker recognition* technologies and methodologies. If the speaker claims to be of a certain identity and the voice is used to verify this claim, this is called *verification* or *authentication*. On the other hand, *identification* is the task of determining an unknown speaker's identity. In a sense *speaker verification* is a 1:1 match where one speaker's voice is matched to one template (also called a "voice print" or "voice model") whereas *speaker identification* is a 1:N match where the voice is compared against N templates.

From a security perspective, identification is different from verification. For example, presenting your passport at border control is a verification process - the agent compares your face to the picture in the document. Conversely, a police officer comparing a sketch of an assailant against a database of previously documented criminals to find the closest match(es) is an identification process.

Speaker verification is usually employed as a "gatekeeper" in order to provide access to a secure system (e.g.: telephone banking). These systems operate with the user's knowledge and typically requires their cooperation. *Speaker identification* systems can also be implemented covertly without the user's knowledge to identify talkers in a discussion, alert automated systems of speaker changes, check if a user is already enrolled in a system, etc.

In forensic applications, it is common to first perform a speaker identification process to create a list of "best matches" and then perform a series of verification processes to determine a conclusive match.

Variants of speaker recognition

Each *speaker recognition* system has two phases: Enrollment and verification. During enrollment, the speaker's voice is recorded and typically a number of features are extracted to form a *voice print*, *template*, or *model*. In the verification phase, a speech sample or "utterance" is compared against a previously created voice print. For identification systems, the utterance is compared against multiple voice prints in order to determine the best match(es) while verification systems compare an utterance against a single voice print. Because of the process involved, verification is faster than identification.

Speaker recognition systems fall into two categories: text-dependent and text-independent.

If the text must be the same for enrollment and verification this is called text-dependent recognition. In a text-dependent system, prompts can either be common across all speakers (e.g.: a common pass phrase) or unique. In addition, the use of shared-secrets (e.g.: passwords and PINs) or knowledge-based information can be employed in order to create a multi-factor authentication scenario.

Text-independent systems are most often used for speaker identification as they require very little if any cooperation by the speaker. In this case the text during enrollment and test is different. In fact, the enrollment may happen without the user's knowledge, as in the case for many forensic applications. As text-independent technologies do not compare what was said at enrollment and verification, verification applications tend to also employ speech recognition to determine what the user is saying at the point of authentication.

Technology

The various technologies used to process and store *voice prints* include frequency estimation, hidden Markov models, Gaussian mixture models, pattern matching algorithms, neural networks, matrix representation,Vector Quantization and decision trees. Some systems also use "anti-speaker" techniques, such as cohort models, and world models.

Ambient noise levels can impede both collection of the initial and subsequent voice samples. Noise reduction algorithms can be employed to improve accuracy, but incorrect application can have the opposite effect. Performance degradation can result from changes in behavioural attributes of the voice and from enrolment using one telephone and verification on another telephone ("cross channel"). Integration with two-factor authentication products is expected to increase. Voice changes due to ageing may impact system performance over time. Some systems adapt the speaker models after each successful verification to capture such long-term changes in the voice, though there is debate regarding the overall security impact imposed by automated adaptation.

Capture of the biometric is seen as non-invasive. The technology traditionally uses existing microphones and voice transmission technology allowing recognition over long distances via ordinary telephones (wired or wireless).

Digitally recorded audio voice identification and analogue recorded voice identification uses electronic measurements as well as critical listening skills that must be applied by a forensic expert in order for the identification to be accurate.

Chapter 6

Voice Stress Analysis

Voice Stress Analysis (VSA) is a controversial lie detection technology. It has been described as pseudoscientific, and there is no known scientific basis for the underlying theory of "microtremors". Federally funded research showed "little validity" in the technique. A study by Virginia State in 2003, at which time the technique was in widespread use, concluded that "Because there have been no independent scientific studies conducted on the reliability of the computer voice analyzer to detect deception, the Board recommends to the Director of the Department of Professional and Occupational Regulation that computer voice analyzer equipment should not be approved in Virginia at this time.", though a number of academic studies are available which call into question the validity of the technique

There is tension between the voice stress analysis community and the polygraph community, due in the main to the fact that the polygraph is heavily regulated and has been subject to numerous detailed scientific studies, while voice stress analysis is largely unregulated and there are few studies (other than by manufacturers and proponents) which show results better than chance.

VSA technology is said to record psychophysiological stress responses that are present in human voice, when a person suffers psychological stress in response to a stimulus (question) and where the consequences of lying may be dire for the subject being 'tested'.

In the Detection Of Deception (DOD) scenario, the voice-stress produced in response to a Relevant Question ("did you do it?") is referred to as psychological stress or 'deceptive stress'. No DOD technology can detect a lie or truth unequivocally. It is the fear of being exposed as lying to the question being posed that produces the 'high stress' voice signature, aka voice graph or voice tracing.

The technique's accuracy remains debated. There are independent research studies that support the use of VSA as a reliable lie detection technology, whilst there are other studies that dispute its reliability.

VSA is distinct from Layered Voice Analysis (LVA). LVA is used to measure many different components of the voice, but is not reliable in the detection of 'deceptive stress'. LVA measures a wide range of emotions, including excitement, confusion, attention and more. LVA is available in many different forms of products, ranging from server based intelligence use systems, to hand-held devices and standard PC software.

The main difference in the method of operation between LVA and VSA is based on the analyzed frequencies ranges: while VSA focuses on the 8–14 Hz range (which is picked up by specialised microphones), LVA uses a wider spectrum range to extract information that is amusing but not particularly relevant to DOD.

Principle and origins

VSA is based on hypothesis that there are infrasonic components of human voice not audible to observers caused by a physiological phenomenon present in muscles called "microtremor". It was discovered in 1957 by British physiologist Olaf Lippold. Further investigation by other researchers explored the possibility of the presence of microtremor in the muscles controlling the voicebox. The experiment was made by attaching electrodes to the cricothyroid muscle and the posterior cricoarytenoid muscle and measuring EMG signals. Detecting microtremor during sustained speech was not deemed possible because the EMG activity changed too rapidly. The experiment was therefore limited to measuring the presence of microtremor in the frequency range of 1 through 20 Hz in sustained vowel phonation, but yielded no positive results. It was concluded that "the electrical energy was randomly distributed throughout the spectrum." The inconclusive research on microtremor in voice production has consequently been used to claim that the phenomenon can be used for creating technology capable of lie detection by detecting microtremor in recorded speech.

Vendors

The original VSA technology was devised by three former US Army personnel. The three, Bell, McQuiston & Ford, developed the PSE 1, an analogue machine. The same three, working under Dektor Counterintelligence and Security Inc., manufactured the PSE 1000 and later the PSE 2000.

The National Institute Of Truth Verification (NITV, West Palm Beach) then produced and marketed an analogue instrument based on the PSE & digitized it in April 1997, based on the McQuiston-Ford algorithm. In the past 10 years VSA has been used primarily in digital applications: Digital Voice Stress Analysis (D-VSA). The primary suppliers in the USA are NITV-CVSA; Dektor Corporation (no relation to the aforementioned Dektor Counterintelligence and Security Inc.), Diogenes-Lantern, and Baker-FVSA. The primary supplier in Malaysia, Singapore, Brunei, India, etc. is the

Australian ITVT Institute (International Truth Verification Technologies) in the form of the Forensic Voice Stress Analyser (FVSA).

The primary use of VSA is in the arena of "Detection Of Deception". As with the polygraph, VSA technology is inert. It has no artificial intelligence component. It is the use of the recorded data as a means for lie detection that remains controversial.

Applications

The purpose of a VSA examination is to determine the truthfulness of responses made by an examinee regarding the subject under investigation. Determinations are made by analyzing and scoring the voice-grams produced by the examinee. Traditional analysis of voice grams was achieved by allocating "percentages of stress" (%) according to the patterns so produced.

High levels of (deceptive) stress indicate that the examinee is deceptive as is the case with polygraph. In respect of VSA, squared voice grams indicates higher stress, whilst 'wave form' or 'domed' signatures indicate less stress.

Questions may be posed to elicit simple "yes" or "no" answers, but can be posed to produce a narrative response. Questions are formulated for each individual being examined to compare situational stress signatures with Control Question and Relevant Question signatures, in order to identify (deceptive) 'stress signatures'.

VSA technology together with validated testing protocols, is designed to protect the innocent and avoid 'false positive' results. VSA is designed to assist any investigation by establishing the veracity of a subject's verbal responses.

Devices used to analyze voice stress are usually used in the presence of the individual under investigation; however, they can also be used without his or her knowledge. Since all that is needed is a voice, a wireless microphone or a tape recording can provide the necessary input signal.

Traditional VSA utilizes the McQuiston-Ford algorithm and this is the technology developed in the USA for the US Defence Agencies and is used by US Law Enforcement agencies.

There are no known physical countermeasures for VSA. Conversely according to Honts *et al.*, the simple use of a 'tack' placed under the tongue of the examinee, to be used as a countermeasure, can reduce the accuracy of polygraph results from 98% to 26%.

Use In law enforcement

A great deal of voice stress testing (VSA) has been conducted. In the United States, most states do not regulate the private use of these devices. However, the CIA and FBI both

use VSA at times, in their own investigations. The technology is currently recognized in 43 states.

Many intelligence agencies as well as private forensic psychophysiologists worldwide utilise VSA in preference to polygraph technology.

The X13-VSA technology was originally developed for military use and is now in use by the Italian State Police and the I.C.A.A. (International Crime Analysis Association), as well by over 150 police agencies and few international airports to screen travelers (X13-VSA PRO Cobra technology).

Methodology and accuracy

The McQuiston-Ford algorithm used for Voice Stress Analysis is reliably accurate. The recorded "micro tremors" in a persons voice are converted via the algorithm into a scorable voice gram. The discrepancy in researched accuracy may result from incorrectly trained or non-trained persons utilizing the technology incorrectly. This is evident by some Polygraphists trying to "test" VSA technology without having received accredited training in the use thereof also by applying Poly Protocols to VSA & vice versa which cannot work.

Recorded cases in Malaysia,Brunei,Singapore,Colombia & recently India have displayed a 100% result with the Australian F V S A which uses a totally new 2006 Developed Algorithm a Scientist & Academic Development

Polygraph-only associations have disputed the accuracy of VSA, although many accredited polygraphists have trained in the use of VSA and use VSA to good effect. The traditional analysis and scoring of voice-grams by means of assigning 'percentages' is time consuming.

In 2002, Clifton Coetzee (Polygraph & VSA Instructor) devised a scoring method for voice grams incorporating the 'UTAH 7 Point' scoring system, as used by modern day polygraphists. Reactive or Responsive patterns are assigned a weighting of +3 to -3.

The use of CQT testing protocols developed by John Reid and Cleve Backster are used for greater reliability of VSA results. It is important that VSA examiners be skilled in the use of enforced, timed pauses between stimulus (question) and response (answer). As in the polygraph situation, the fight or flight response has onset and conclusion delays, which must be considered by examiners to achieve reliable results.

The American Polygraph Association's website lists conclusions from multiple studies into the accuracy of voice stress analysis as a means of detecting the subject's truthfulness. Some researchers or polygraph professionals cast doubt on the validity of the results of such tests; many describe the results as no better than chance.

A study from the U.S Department of Justice showed that VSA performed with a 50% accuracy rate. 2008 W.Carolina University Paper 140 - Eng 101 shows that while VSA works that the FFT & Mcquiston Algorthims do not totally capture (FFT is used for Polygraph also) EMD is more accurate- Authors Prof. Tay , Asst Professor Adams etc.

Chapter 7

Speech Recognition

The display of the Speech Recognition screensaver on a PC, in which the character responds to questions, e.g. "Where are you?" or statements, e.g. "Hello."

Speech recognition (also known as **automatic speech recognition** or **computer speech recognition**) converts spoken words to text. The term "voice recognition" is sometimes used to refer to recognition systems that must be trained to a particular speaker—as is the case for most desktop recognition software. Recognizing the speaker can simplify the task of translating speech.

Speech recognition is a broader solution which refers to technology that can recognize speech without being targeted at single speaker—such as a call system that can recognize arbitrary voices.

Speech recognition applications include voice user interfaces such as voice dialing (e.g., "Call home"), call routing (e.g., "I would like to make a collect call"), domotic appliance control, search (e.g., find a podcast where particular words were spoken), simple data entry (e.g., entering a credit card number), preparation of structured documents (e.g., a radiology report), speech-to-text processing (e.g., word processors or emails), and aircraft (usually termed Direct Voice Input).

History

The first speech recognizer appeared in 1952 and consisted of a device for the recognition of single spoken digits Another early device was the IBM Shoebox, exhibited at the 1964 New York World's Fair.

One of the most notable domains for the commercial application of speech recognition in the United States has been health care and in particular the work of the medical transcriptionist (MT). According to industry experts, at its inception, speech recognition (SR) was sold as a way to completely eliminate transcription rather than make the transcription process more efficient, hence it was not accepted. It was also the case that SR at that time was often technically deficient. Additionally, to be used effectively, it required changes to the ways physicians worked and documented clinical encounters, which many if not all were reluctant to do. The biggest limitation to speech recognition automating transcription, however, is seen as the software. The nature of narrative dictation is highly interpretive and often requires judgment that may be provided by a real human but not yet by an automated system. Another limitation has been the extensive amount of time required by the user and/or system provider to train the software.

A distinction in ASR is often made between "artificial syntax systems" which are usually domain-specific and "natural language processing" which is usually language-specific. Each of these types of application presents its own particular goals and challenges.

Applications

Health care

In the health care domain, even in the wake of improving speech recognition technologies, medical transcriptionists (MTs) have not yet become obsolete. The services provided may be redistributed rather than replaced.

Speech recognition can be implemented in front-end or back-end of the medical documentation process.

Front-End SR is where the provider dictates into a speech-recognition engine, the recognized words are displayed right after they are spoken, and the dictator is responsible for editing and signing off on the document. It never goes through an MT/editor.

Back-End SR or Deferred SR is where the provider dictates into a digital dictation system, and the voice is routed through a speech-recognition machine and the recognized draft document is routed along with the original voice file to the MT/editor, who edits the draft and finalizes the report. Deferred SR is being widely used in the industry currently.

Many Electronic Medical Records (EMR) applications can be more effective and may be performed more easily when deployed in conjunction with a speech-recognition engine.

Searches, queries, and form filling may all be faster to perform by voice than by using a keyboard.

Healthcare solutions are usually very state-specific, however some companies adjust their solutions to the needs of concreete markets (p.e. Speech Technology Center in Russia has a Finnish partner Vitim OY with a "Terve Elama" project).

Military

High-performance fighter aircraft

Substantial efforts have been devoted in the last decade to the test and evaluation of speech recognition in fighter aircraft. Of particular note is the U.S. program in speech recognition for the Advanced Fighter Technology Integration (AFTI)/F-16 aircraft (F-16 VISTA), and a program in France installing speech recognition systems on Mirage aircraft, and also programs in the UK dealing with a variety of aircraft platforms. In these programs, speech recognizers have been operated successfully in fighter aircraft, with applications including: setting radio frequencies, commanding an autopilot system, setting steer-point coordinates and weapons release parameters, and controlling flight displays.

Working with Swedish pilots flying in the JAS-39 Gripen cockpit, Englund (2004) found recognition deteriorated with increasing G-loads. It was also concluded that adaptation greatly improved the results in all cases and introducing models for breathing was shown to improve recognition scores significantly. Contrary to what might be expected, no effects of the broken English of the speakers were found. It was evident that spontaneous speech caused problems for the recognizer, as could be expected. A restricted vocabulary, and above all, a proper syntax, could thus be expected to improve recognition accuracy substantially.

The Eurofighter Typhoon currently in service with the UK RAF employs a speaker-dependent system, i.e. it requires each pilot to create a template. The system is not used for any safety critical or weapon critical tasks, such as weapon release or lowering of the undercarriage, but is used for a wide range of other cockpit functions. Voice commands are confirmed by visual and/or aural feedback. The system is seen as a major design feature in the reduction of pilot workload, and even allows the pilot to assign targets to himself with two simple voice commands or to any of his wingmen with only five commands.

Speaker independent systems are also being developed and are in testing for the F35 Lightning II (JSF) and the Alenia Aermacchi M-346 Master lead-in fighter trainer. These systems have produced word accuracies in excess of 98%.

Helicopters

The problems of achieving high recognition accuracy under stress and noise pertain strongly to the helicopter environment as well as to the fighter environment. The acoustic noise problem is actually more severe in the helicopter environment, not only because of the high noise levels but also because the helicopter pilot generally does not wear a facemask, which would reduce acoustic noise in the microphone. Substantial test and evaluation programs have been carried out in the past decade in speech recognition systems applications in helicopters, notably by the U.S. Army Avionics Research and Development Activity (AVRADA) and by the Royal Aerospace Establishment (RAE) in the UK. Work in France has included speech recognition in the Puma helicopter. There has also been much useful work in Canada. Results have been encouraging, and voice applications have included: control of communication radios; setting of navigation systems; and control of an automated target handover system.

As in fighter applications, the overriding issue for voice in helicopters is the impact on pilot effectiveness. Encouraging results are reported for the AVRADA tests, although these represent only a feasibility demonstration in a test environment. Much remains to be done both in speech recognition and in overall speech recognition technology, in order to consistently achieve performance improvements in operational settings.

Battle management

Battle Management command centres generally require rapid access to and control of large, rapidly changing information databases. Commanders and system operators need to query these databases as conveniently as possible, in an eyes-busy environment where much of the information is presented in a display format. Human-machine interaction by voice has the potential to be very useful in these environments. A number of efforts have been undertaken to interface commercially available isolated-word recognizers into battle management environments. In one feasibility study speech recognition equipment was tested in conjunction with an integrated information display for naval battle management applications. Users were very optimistic about the potential of the system, although capabilities were limited.

Speech understanding programs sponsored by the Defense Advanced Research Projects Agency (DARPA) in the U.S. has focused on this problem of natural speech interface. Speech recognition efforts have focused on a database of continuous speech recognition (CSR), large-vocabulary speech which is designed to be representative of the naval resource management task. Significant advances in the state-of-the-art in CSR have been achieved, and current efforts are focused on integrating speech recognition and natural language processing to allow spoken language interaction with a naval resource management system.

Training air traffic controllers

Training for air traffic controllers (ATC) represents an excellent application for speech recognition systems. Many ATC training systems currently require a person to act as a "pseudo-pilot", engaging in a voice dialog with the trainee controller, which simulates the dialog which the controller would have to conduct with pilots in a real ATC situation. Speech recognition and synthesis techniques offer the potential to eliminate the need for a person to act as pseudo-pilot, thus reducing training and support personnel. In theory, Air controller tasks are also characterized by highly structured speech as the primary output of the controller, hence reducing the difficulty of the speech recognition task should be possible. In practice this is rarely the case. The FAA document 7110.65 details the phrases that should be used by air traffic controllers. While this document gives less than 150 examples of such phrases, the number of phrases supported by one of the simulation vendors speech recognition systems is in excess of 500,000.

The USAF, USMC, US Army,US Navy and FAA as well as a number of international ATC training organizations such as the Royal Australian Air Force and Civil Aviation Authorities in Italy, Brazil, Canada are currently using ATC simulators with speech recognition from a number of different vendors.

Telephony and other domains

ASR in the field of telephony is now commonplace and in the field of computer gaming and simulation is becoming more widespread. Despite the high level of integration with word processing in general personal computing, however, ASR in the field of document production has not seen the expected increases in use.

The improvement of mobile processor speeds made feasible the speech-enabled Symbian and Windows Mobile Smartphones. Speech is used mostly as a part of User Interface, for creating pre-defined or custom speech commands. Leading software vendors in this field are: Microsoft Corporation (Microsoft Voice Command), Digital Syphon (Sonic Extractor), Nuance Communications (Nuance Voice Control), Speech Technology Center, Vito Technology (VITO Voice2Go), Speereo Software (Speereo Voice Translator and SVOX.

Further applications

- Automatic translation;
- Automotive speech recognition (e.g., Ford Sync);
- Telematics (e.g. vehicle Navigation Systems);
- Court reporting (Realtime Voice Writing);
- Hands-free computing: voice command recognition computer user interface;
- Home automation;
- Interactive voice response;
- Mobile telephony, including mobile email;
- Multimodal interaction;

- Pronunciation evaluation in computer-aided language learning applications;
- Robotics;
- Video games, with Tom Clancy's EndWar and Lifeline as working examples;
- Transcription (digital speech-to-text);
- Speech-to-text (transcription of speech into mobile text messages);

Performance

The performance of speech recognition systems is usually specified in terms of accuracy and speed. Accuracy is usually rated with word error rate (WER), whereas speed is measured with the real time factor. Other measures of accuracy include Single Word Error Rate (SWER) and Command Success Rate (CSR).

In 1982, Kurzweil Applied Intelligence and Dragon Systems released speech recognition products. By 1985, Kurzweil's software had a vocabulary of 1,000 words—if uttered one word at a time. Two years later, in 1987, its lexicon reached 20,000 words, entering the realm of human vocabularies, which range from 10,000 to 150,000 words. But recognition accuracy was only 10% in 1993. Two years later, the error rate crossed below 50%. Dragon Systems released "Naturally Speaking" in 1997 which recognized normal human speech. Progress mainly came from improved computer performance and larger source text databases. The Brown Corpus was the first major database available, containing several million words. In 2006, Google published a trillion-word corpus, while Carnegie Mellon University researchers found no significant increase in recognition accuracy.

Algorithms

Both acoustic modeling and language modeling are important parts of modern statistically-based speech recognition algorithms. Hidden Markov models (HMMs) are widely used in many systems. Language modeling has many other applications such as smart keyboard and document classification.

Hidden Markov models

Modern general-purpose speech recognition systems are based on Hidden Markov Models. These are statistical models which output a sequence of symbols or quantities. HMMs are used in speech recognition because a speech signal can be viewed as a piecewise stationary signal or a short-time stationary signal. In a short-time (e.g., 10 milliseconds)), speech can be approximated as a stationary process. Speech can be thought of as a Markov model for many stochastic purposes.

Another reason why HMMs are popular is because they can be trained automatically and are simple and computationally feasible to use. In speech recognition, the hidden Markov model would output a sequence of n-dimensional real-valued vectors (with n being a small integer, such as 10), outputting one of these every 10 milliseconds. The vectors would consist of cepstral coefficients, which are obtained by taking a Fourier transform

of a short time window of speech and decorrelating the spectrum using a cosine transform, then taking the first (most significant) coefficients. The hidden Markov model will tend to have in each state a statistical distribution that is a mixture of diagonal covariance Gaussians which will give a likelihood for each observed vector. Each word, or (for more general speech recognition systems), each phoneme, will have a different output distribution; a hidden Markov model for a sequence of words or phonemes is made by concatenating the individual trained hidden Markov models for the separate words and phonemes.

Described above are the core elements of the most common, HMM-based approach to speech recognition. Modern speech recognition systems use various combinations of a number of standard techniques in order to improve results over the basic approach described above. A typical large-vocabulary system would need context dependency for the phonemes (so phonemes with different left and right context have different realizations as HMM states); it would use cepstral normalization to normalize for different speaker and recording conditions; for further speaker normalization it might use vocal tract length normalization (VTLN) for male-female normalization and maximum likelihood linear regression (MLLR) for more general speaker adaptation. The features would have so-called delta and delta-delta coefficients to capture speech dynamics and in addition might use heteroscedastic linear discriminant analysis (HLDA); or might skip the delta and delta-delta coefficients and use splicing and an LDA-based projection followed perhaps by heteroscedastic linear discriminant analysis or a global semitied covariance transform (also known as maximum likelihood linear transform, or MLLT). Many systems use so-called discriminative training techniques which dispense with a purely statistical approach to HMM parameter estimation and instead optimize some classification-related measure of the training data. Examples are maximum mutual information (MMI), minimum classification error (MCE) and minimum phone error (MPE).

Decoding of the speech (the term for what happens when the system is presented with a new utterance and must compute the most likely source sentence) would probably use the Viterbi algorithm to find the best path, and here there is a choice between dynamically creating a combination hidden Markov model which includes both the acoustic and language model information, or combining it statically beforehand (the finite state transducer, or FST, approach).

A possible improvement to decoding is to keep a set of good candidates instead of just keeping the best candidate, and to use a better scoring function (rescoring) to rate these good candidates so that we may pick the best one according to this refined score. The set of candidates can either be kept as a list (the N-best list approach) or as a subset of the models (a lattice). Rescoring is usually done by trying to minimize the Bayes risk (or an approximation thereof): instead of taking the source sentence with maximal probability, we try to take the sentence which minimizes the expectancy of a given loss function with regards to all possible transcriptions (ie. we take the sentence which minimizes the average distance to other possible sentences weighted by their estimated probability). The loss function is usually the Levenshtein distance, though it can be different distances for

specific tasks; the set of possible transcriptions is of course pruned to maintain tractability. Efficient algorithms have been devised to rescore lattices represented as weighted finite state transducers with edit distances represented themselves as a finite state transducer verifying certain assumptions.

Dynamic time warping (DTW)-based speech recognition

Dynamic time warping is an approach that was historically used for speech recognition but has now largely been displaced by the more successful HMM-based approach. Dynamic time warping is an algorithm for measuring similarity between two sequences which may vary in time or speed. For instance, similarities in walking patterns would be detected, even if in one video the person was walking slowly and if in another they were walking more quickly, or even if there were accelerations and decelerations during the course of one observation. DTW has been applied to video, audio, and graphics – indeed, any data which can be turned into a linear representation can be analyzed with DTW.

A well known application has been automatic speech recognition, to cope with different speaking speeds. In general, it is a method that allows a computer to find an optimal match between two given sequences (e.g. time series) with certain restrictions, i.e. the sequences are "warped" non-linearly to match each other. This sequence alignment method is often used in the context of hidden Markov models.

People with disabilities

People with disabilities can benefit from speech recognition programs. Speech recognition is especially useful for people who have difficulty using their hands, ranging from mild repetitive stress injuries to involved disabilities that preclude using conventional computer input devices. In fact, people who used the keyboard a lot and developed RSI became an urgent early market for speech recognition. Speech recognition is used in deaf telephony, such as voicemail to text, relay services, and captioned telephone. Individuals with learning disabilities who have problems with thought-to-paper communication (essentially they think of an idea but it is processed incorrectly causing it to end up differently on paper) can benefit from the software.

Current research and funding

Measuring progress in speech recognition performance is difficult and controversial. Some speech recognition tasks are much more difficult than others. Word error rates on some tasks are less than 1%. On others they can be as high as 50%. Sometimes it even appears that performance is going backwards as researchers undertake harder tasks that have higher error rates.

Because progress is slow and is difficult to measure, there is some perception that performance has plateaued and that funding has dried up or shifted priorities. Such perceptions are not new. In 1969, John Pierce wrote an open letter that did cause much funding to dry up for several years. In 1993 there was a strong feeling that performance

had plateaued and there were workshops dedicated to the issue. However, in the 1990s funding continued more or less uninterrupted and performance continued slowly but steadily to improve.

For the past thirty years, speech recognition research has been characterized by the steady accumulation of small incremental improvements. There has also been a trend continually to change focus to more difficult tasks due both to progress in speech recognition performance and to the availability of faster computers. In particular, this shifting to more difficult tasks has characterized DARPA funding of speech recognition since the 1980s. In the last decade it has continued with the EARS project, which undertook recognition of Mandarin and Arabic in addition to English, and the GALE project, which focused solely on Mandarin and Arabic and required translation simultaneously with speech recognition.

Commercial research and other academic research also continue to focus on increasingly difficult problems. One key area is to improve robustness of speech recognition performance, not just robustness against noise but robustness against any condition that causes a major degradation in performance. Another key area of research is focused on an opportunity rather than a problem. This research attempts to take advantage of the fact that in many applications there is a large quantity of speech data available, up to millions of hours. It is too expensive to have humans transcribe such large quantities of speech, so the research focus is on developing new methods of machine learning that can effectively utilize large quantities of unlabeled data. Another area of research is better understanding of human capabilities and to use this understanding to improve machine recognition performance.

Chapter 8

Speech Synthesis

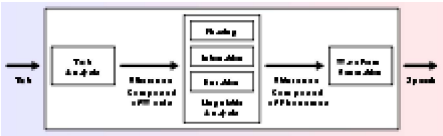
Stephen Hawking is one of the most famous people using speech synthesis to communicate

Speech synthesis is the artificial production of human speech. A computer system used for this purpose is called a **speech synthesizer**, and can be implemented in software or hardware. A **text-to-speech (TTS)** system converts normal language text into speech; other systems render symbolic linguistic representations like phonetic transcriptions into speech.

Synthesized speech can be created by concatenating pieces of recorded speech that are stored in a database. Systems differ in the size of the stored speech units; a system that stores phones or diphones provides the largest output range, but may lack clarity. For specific usage domains, the storage of entire words or sentences allows for high-quality output. Alternatively, a synthesizer can incorporate a model of the vocal tract and other human voice characteristics to create a completely "synthetic" voice output.

The quality of a speech synthesizer is judged by its similarity to the human voice and by its ability to be understood. An intelligible text-to-speech program allows people with visual impairments or reading disabilities to listen to written works on a home computer. Many computer operating systems have included speech synthesizers since the early 1980s.

Overview of text processing



Overview of a typical TTS system

A text-to-speech system (or "engine") is composed of two parts: a front-end and a back-end. The front-end has two major tasks. First, it converts raw text containing symbols like numbers and abbreviations into the equivalent of written-out words. This process is often called *text normalization*, *pre-processing*, or *tokenization*. The front-end then assigns phonetic transcriptions to each word, and divides and marks the text into prosodic units, like phrases, clauses, and sentences. The process of assigning phonetic transcriptions to words is called *text-to-phoneme* or *grapheme-to-phoneme* conversion. Phonetic transcriptions and prosody information together make up the symbolic linguistic representation that is output by the front-end. The back-end—often referred to as the *synthesizer*—then converts the symbolic linguistic representation into sound. In certain systems, this part includes the computation of the *target prosody* (pitch contour, phoneme durations), which is then imposed on the output speech.

History

Long before electronic signal processing was invented, there were those who tried to build machines to create human speech. Some early legends of the existence of "speaking heads" involved Gerbert of Aurillac (d. 1003 AD), Albertus Magnus (1198–1280), and Roger Bacon (1214–1294).

In 1779, the Danish scientist Christian Kratzenstein, working at the Russian Academy of Sciences, built models of the human vocal tract that could produce the five long vowel sounds (in International Phonetic Alphabet notation, they are [a:], [e:], [i:], [o:] and [u:]). This was followed by the bellows-operated "acoustic-mechanical speech machine" by Wolfgang von Kempelen of Vienna, Austria, described in a 1791 paper. This machine added models of the tongue and lips, enabling it to produce consonants as well as vowels. In 1837, Charles Wheatstone produced a "speaking machine" based on von Kempelen's design, and in 1857, M. Faber built the "Euphonia". Wheatstone's design was resurrected in 1923 by Paget.

In the 1930s, Bell Labs developed the VOCODER, a keyboard-operated electronic speech analyzer and synthesizer that was said to be clearly intelligible. Homer Dudley refined this device into the VODER, which he exhibited at the 1939 New York World's Fair.

The Pattern playback was built by Dr. Franklin S. Cooper and his colleagues at Haskins Laboratories in the late 1940s and completed in 1950. There were several different versions of this hardware device but only one currently survives. The machine converts

pictures of the acoustic patterns of speech in the form of a spectrogram back into sound. Using this device, Alvin Liberman and colleagues were able to discover acoustic cues for the perception of phonetic segments (consonants and vowels).

Dominant systems in the 1980s and 1990s were the MITalk system, based largely on the work of Dennis Klatt at MIT, and the Bell Labs system; the latter was one of the first multilingual language-independent systems, making extensive use of Natural Language Processing methods.

Early electronic speech synthesizers sounded robotic and were often barely intelligible. The quality of synthesized speech has steadily improved, but output from contemporary speech synthesis systems is still clearly distinguishable from actual human speech.

As the cost-performance ratio causes speech synthesizers to become cheaper and more accessible to the people, more people will benefit from the use of text-to-speech programs.

Electronic devices

The first computer-based speech synthesis systems were created in the late 1950s, and the first complete text-to-speech system was completed in 1968. In 1961, physicist John Larry Kelly, Jr and colleague Louis Gerstman used an IBM 704 computer to synthesize speech, an event among the most prominent in the history of Bell Labs. Kelly's voice recorder synthesizer (vocoder) recreated the song "Daisy Bell", with musical accompaniment from Max Mathews. Coincidentally, Arthur C. Clarke was visiting his friend and colleague John Pierce at the Bell Labs Murray Hill facility. Clarke was so impressed by the demonstration that he used it in the climactic scene of his screenplay for his novel *2001: A Space Odyssey*, where the HAL 9000 computer sings the same song as it is being put to sleep by astronaut Dave Bowman. Despite the success of purely electronic speech synthesis, research is still being conducted into mechanical speech synthesizers.

Handheld electronics featuring speech synthesis began emerging in the 1970s. One of the first was the Telesensory Systems Inc. (TSI) *Speech+* portable calculator for the blind in 1976. Other devices were produced primarily for educational purposes, such as Speak & Spell, produced by Texas Instruments in 1978. Fidelity released a speaking version of its electronic chess computer in 1979. The first video game to feature speech synthesis was the shoot 'em up arcade game, *Stratovox*, from Sun Electronics and the arcade version of Bezerk both released in 1980. The first multi-player game using voice synthesis was *Milton* from Milton Bradley Company, which produced the device in 1980.

Synthesizer technologies

The most important qualities of a speech synthesis system are *naturalness* and *intelligibility*. Naturalness describes how closely the output sounds like human speech, while intelligibility is the ease with which the output is understood. The ideal speech

synthesizer is both natural and intelligible. Speech synthesis systems usually try to maximize both characteristics.

The two primary technologies for generating synthetic speech waveforms are *concatenative synthesis* and *formant synthesis*. Each technology has strengths and weaknesses, and the intended uses of a synthesis system will typically determine which approach is used.

Concatenative synthesis

Concatenative synthesis is based on the concatenation (or stringing together) of segments of recorded speech. Generally, concatenative synthesis produces the most natural-sounding synthesized speech. However, differences between natural variations in speech and the nature of the automated techniques for segmenting the waveforms sometimes result in audible glitches in the output. There are three main sub-types of concatenative synthesis.

Unit selection synthesis

Unit selection synthesis uses large databases of recorded speech. During database creation, each recorded utterance is segmented into some or all of the following: individual phones, diphones, half-phones, syllables, morphemes, words, phrases, and sentences. Typically, the division into segments is done using a specially modified speech recognizer set to a "forced alignment" mode with some manual correction afterward, using visual representations such as the waveform and spectrogram. An index of the units in the speech database is then created based on the segmentation and acoustic parameters like the fundamental frequency (pitch), duration, position in the syllable, and neighboring phones. At runtime, the desired target utterance is created by determining the best chain of candidate units from the database (unit selection). This process is typically achieved using a specially weighted decision tree.

Unit selection provides the greatest naturalness, because it applies only a small amount of digital signal processing (DSP) to the recorded speech. DSP often makes recorded speech sound less natural, although some systems use a small amount of signal processing at the point of concatenation to smooth the waveform. The output from the best unit-selection systems is often indistinguishable from real human voices, especially in contexts for which the TTS system has been tuned. However, maximum naturalness typically require unit-selection speech databases to be very large, in some systems ranging into the gigabytes of recorded data, representing dozens of hours of speech. Also, unit selection algorithms have been known to select segments from a place that results in less than ideal synthesis (e.g. minor words become unclear) even when a better choice exists in the database.

Diphone synthesis

Diphone synthesis uses a minimal speech database containing all the diphones (sound-to-sound transitions) occurring in a language. The number of diphones depends on the phonotactics of the language: for example, Spanish has about 800 diphones, and German about 2500. In diphone synthesis, only one example of each diphone is contained in the speech database. At runtime, the target prosody of a sentence is superimposed on these minimal units by means of digital signal processing techniques such as linear predictive coding, PSOLA or MBROLA. The quality of the resulting speech is generally worse than that of unit-selection systems, but more natural-sounding than the output of formant synthesizers. Diphone synthesis suffers from the sonic glitches of concatenative synthesis and the robotic-sounding nature of formant synthesis, and has few of the advantages of either approach other than small size. As such, its use in commercial applications is declining, although it continues to be used in research because there are a number of freely available software implementations.

Domain-specific synthesis

Domain-specific synthesis concatenates prerecorded words and phrases to create complete utterances. It is used in applications where the variety of texts the system will output is limited to a particular domain, like transit schedule announcements or weather reports. The technology is very simple to implement, and has been in commercial use for a long time, in devices like talking clocks and calculators. The level of naturalness of these systems can be very high because the variety of sentence types is limited, and they closely match the prosody and intonation of the original recordings.

Because these systems are limited by the words and phrases in their databases, they are not general-purpose and can only synthesize the combinations of words and phrases with which they have been preprogrammed. The blending of words within naturally spoken language however can still cause problems unless the many variations are taken into account. This alternation cannot be reproduced by a simple word-concatenation system, which would require additional complexity to be context-sensitive.

Formant synthesis

Formant synthesis does not use human speech samples at runtime. Instead, the synthesized speech output is created using additive synthesis and an acoustic model (physical modelling synthesis). Parameters such as fundamental frequency, voicing, and noise levels are varied over time to create a waveform of artificial speech. This method is sometimes called *rules-based synthesis*; however, many concatenative systems also have rules-based components. Many systems based on formant synthesis technology generate artificial, robotic-sounding speech that would never be mistaken for human speech. However, maximum naturalness is not always the goal of a speech synthesis system, and formant synthesis systems have advantages over concatenative systems. Formant-synthesized speech can be reliably intelligible, even at very high speeds, avoiding the acoustic glitches that commonly plague concatenative systems. High-speed synthesized

speech is used by the visually impaired to quickly navigate computers using a screen reader. Formant synthesizers are usually smaller programs than concatenative systems because they do not have a database of speech samples. They can therefore be used in embedded systems, where memory and microprocessor power are especially limited. Because formant-based systems have complete control of all aspects of the output speech, a wide variety of prosodies and intonations can be output, conveying not just questions and statements, but a variety of emotions and tones of voice.

Examples of non-real-time but highly accurate intonation control in formant synthesis include the work done in the late 1970s for the Texas Instruments toy Speak & Spell, and in the early 1980s Sega arcade machines and in many Atari, Inc. arcade games using the TMS5220 LPC Chips. Creating proper intonation for these projects was painstaking, and the results have yet to be matched by real-time text-to-speech interfaces.

Articulatory synthesis

Articulatory synthesis refers to computational techniques for synthesizing speech based on models of the human vocal tract and the articulation processes occurring there. The first articulatory synthesizer regularly used for laboratory experiments was developed at Haskins Laboratories in the mid-1970s by Philip Rubin, Tom Baer, and Paul Mermelstein. This synthesizer, known as ASY, was based on vocal tract models developed at Bell Laboratories in the 1960s and 1970s by Paul Mermelstein, Cecil Coker, and colleagues.

Until recently, articulatory synthesis models have not been incorporated into commercial speech synthesis systems. A notable exception is the NeXT-based system originally developed and marketed by Trillium Sound Research, a spin-off company of the University of Calgary, where much of the original research was conducted. Following the demise of the various incarnations of NeXT (started by Steve Jobs in the late 1980s and merged with Apple Computer in 1997), the Trillium software was published under the GNU General Public License, with work continuing as gnuspeech. The system, first marketed in 1994, provides full articulatory-based text-to-speech conversion using a waveguide or transmission-line analog of the human oral and nasal tracts controlled by Carré's "distinctive region model".

HMM-based synthesis

HMM-based synthesis is a synthesis method based on hidden Markov models, also called Statistical Parametric Synthesis. In this system, the frequency spectrum (vocal tract), fundamental frequency (vocal source), and duration (prosody) of speech are modeled simultaneously by HMMs. Speech waveforms are generated from HMMs themselves based on the maximum likelihood criterion.

Sinewave synthesis

Sinewave synthesis is a technique for synthesizing speech by replacing the formants (main bands of energy) with pure tone whistles.

Challenges

Text normalization challenges

The process of normalizing text is rarely straightforward. Texts are full of heteronyms, numbers, and abbreviations that all require expansion into a phonetic representation. There are many spellings in English which are pronounced differently based on context. For example, "My latest project is to learn how to better project my voice" contains two pronunciations of "project".

Most text-to-speech (TTS) systems do not generate semantic representations of their input texts, as processes for doing so are not reliable, well understood, or computationally effective. As a result, various heuristic techniques are used to guess the proper way to disambiguate homographs, like examining neighboring words and using statistics about frequency of occurrence.

Recently TTS systems have begun to use HMMs (discussed above) to generate "parts of speech" to aid in disambiguating homographs. This technique is quite successful for many cases such as whether "read" should be pronounced as "red" implying past tense, or as "reed" implying present tense. Typical error rates when using HMMs in this fashion are usually below five percent. These techniques also work well for most European languages, although access to required training corpora is frequently difficult in these languages.

Deciding how to convert numbers is another problem that TTS systems have to address. It is a simple programming challenge to convert a number into words (at least in English), like "1325" becoming "one thousand three hundred twenty-five." However, numbers occur in many different contexts; "1325" may also be read as "one three two five", "thirteen twenty-five" or "thirteen hundred and twenty five". A TTS system can often infer how to expand a number based on surrounding words, numbers, and punctuation, and sometimes the system provides a way to specify the context if it is ambiguous. Roman numerals can also be read differently depending on context. For example "Henry VIII" reads as "Henry the Eighth", while "Chapter VIII" reads as "Chapter Eight".

Similarly, abbreviations can be ambiguous. For example, the abbreviation "in" for "inches" must be differentiated from the word "in", and the address "12 St John St." uses the same abbreviation for both "Saint" and "Street". TTS systems with intelligent front ends can make educated guesses about ambiguous abbreviations, while others provide the same result in all cases, resulting in nonsensical (and sometimes comical) outputs.

Text-to-phoneme challenges

Speech synthesis systems use two basic approaches to determine the pronunciation of a word based on its spelling, a process which is often called text-to-phoneme or grapheme-to-phoneme conversion (phoneme is the term used by linguists to describe distinctive sounds in a language). The simplest approach to text-to-phoneme conversion is the dictionary-based approach, where a large dictionary containing all the words of a language and their correct pronunciations is stored by the program. Determining the correct pronunciation of each word is a matter of looking up each word in the dictionary and replacing the spelling with the pronunciation specified in the dictionary. The other approach is rule-based, in which pronunciation rules are applied to words to determine their pronunciations based on their spellings. This is similar to the "sounding out", or synthetic phonics, approach to learning reading.

Each approach has advantages and drawbacks. The dictionary-based approach is quick and accurate, but completely fails if it is given a word which is not in its dictionary. As dictionary size grows, so too does the memory space requirements of the synthesis system. On the other hand, the rule-based approach works on any input, but the complexity of the rules grows substantially as the system takes into account irregular spellings or pronunciations. (Consider that the word "of" is very common in English, yet is the only word in which the letter "f" is pronounced [v].) As a result, nearly all speech synthesis systems use a combination of these approaches.

Languages with a phonemic orthography have a very regular writing system, and the prediction of the pronunciation of words based on their spellings is quite successful. Speech synthesis systems for such languages often use the rule-based method extensively, resorting to dictionaries only for those few words, like foreign names and borrowings, whose pronunciations are not obvious from their spellings. On the other hand, speech synthesis systems for languages like English, which have extremely irregular spelling systems, are more likely to rely on dictionaries, and to use rule-based methods only for unusual words, or words that aren't in their dictionaries.

Evaluation challenges

The consistent evaluation of speech synthesis systems may be difficult because of a lack of universally agreed objective evaluation criteria. Different organizations often use different speech data. The quality of speech synthesis systems also depends to a large degree on the quality of the production technique (which may involve analogue or digital recording) and on the facilities used to replay the speech. Evaluating speech synthesis systems has therefore often been compromised by differences between production techniques and replay facilities.

Recently, however, some researchers have started to evaluate speech synthesis systems using a common speech dataset.

Prosodics and emotional content

A recent study reported in the journal "**Speech Communication**" by Amy Drahota and colleagues at the University of Portsmouth, UK, reported that listeners to voice recordings could determine, at better than chance levels, whether or not the speaker was smiling. It was suggested that identification of the vocal features which signal emotional content may be used to help make synthesized speech sound more natural.

Dedicated hardware

- Votrax
  - SC-01A (analog formant)
  - SC-02 / SSI-263 / "Artic 263"
- General Instruments SPO256-AL2 (CTS256A-AL2)
- Magnevation SpeakJet
- Savage Innovations SoundGin
- National Semiconductor DT1050 Digitalker (Mozer)
- Silicon Systems SSI 263 (analog formant)
- Texas Instruments LPC Speech Chips
  - TMS5110A
  - TMS5200
- Oki Semiconductor
  - ML22825 (ADPCM)
  - ML22573 (HQADPCM)
- Toshiba T6721A
- Philips / Signetics
  - Mullard MEA8000
  - PCF8200
- TextSpeak Embedded TTS Modules

Computer operating systems or outlets with speech synthesis

Atari

Arguably, the first speech system integrated into an operating system was the 1400XL/1450XL personal computers designed by Atari, Inc. using the Votrax SC01 chip in 1983. The 1400XL/1450XL computers used a Finite State Machine to enable World English Spelling text-to-speech synthesis. Unfortunately, the 1400XL/1450XL personal computers never shipped in quantity.

The Atari ST computers were sold with "stspeech.tos" on floppy disk.

Apple

The first speech system integrated into an operating system that shipped in quantity was Apple Computer's MacInTalk in 1984. Since the 1980s Macintosh Computers offered

text to speech capabilities through The MacinTalk software. In the early 1990s Apple expanded its capabilities offering system wide text-to-speech support. With the introduction of faster PowerPC-based computers they included higher quality voice sampling. Apple also introduced speech recognition into its systems which provided a fluid command set. More recently, Apple has added sample-based voices. Starting as a curiosity, the speech system of Apple Macintosh has evolved into a fully-supported program, PlainTalk, for people with vision problems. VoiceOver was for the first time featured in Mac OS X Tiger (10.4). During 10.4 (Tiger) & first releases of 10.5 (Leopard) there was only one standard voice shipping with Mac OS X. Starting with 10.6 (Snow Leopard), the user can choose out of a wide range list of multiple voices. VoiceOver voices feature the taking of realistic-sounding breaths between sentences, as well as improved clarity at high read rates over PlainTalk. Mac OS X also includes say, a command-line based application that converts text to audible speech. The AppleScript Standard Additions includes a say verb that allows a script to use any of the installed voices and to control the pitch, speaking rate and modulation of the spoken text.

The Apple iOS operating system used on the iPhone, iPad and iPod Touch uses VoiceOver speech synthesis for accessibility. Some third party applications also provide speech synthesis to facilitate navigating, reading web pages or translating text.

AmigaOS

The second operating system with advanced speech synthesis capabilities was AmigaOS, introduced in 1985. The voice synthesis was licensed by Commodore International from a third-party software house (Don't Ask Software, now Softvoice, Inc.) and it featured a complete system of voice emulation, with both male and female voices and "stress" indicator markers, made possible by advanced features of the Amiga hardware audio chipset. It was divided into a narrator device and a translator library. Amiga Speak Handler featured a text-to-speech translator. AmigaOS considered speech synthesis a virtual hardware device, so the user could even redirect console output to it. Some Amiga programs, such as word processors, made extensive use of the speech system.

Microsoft Windows

Modern Windows desktop systems can use SAPI 4 and SAPI 5 components to support speech synthesis and speech recognition. SAPI 4.0 was available as an optional add-on for Windows 95 and Windows 98. Windows 2000 added Narrator, a text–to–speech utility for people who have visual handicaps. Third-party programs such as CoolSpeech can perform various text-to-speech tasks such as reading text aloud from a specified website, email account, text document, the Windows clipboard, the user's keyboard typing, etc. Not all programs can use speech synthesis directly. Some programs can use plug-ins, extensions or add-ons to read text aloud. Third-party programs are available that can read text from the system clipboard.

Microsoft Speech Server is a server-based package for voice synthesis and recognition. It is designed for network use with web applications and call centers.

Text-to-Speech (**TTS**) refers to the ability of computers to read text aloud. A **TTS Engine** converts written text to a phonemic representation, then converts the phonemic representation to waveforms that can be output as sound. TTS engines with different languages, dialects and specialized vocabularies are available through third-party publishers.

Android

Version 1.6 of Android added support for speech synthesis (TTS).

Internet

Currently, there are a number of applications, plugins and gadgets that can read messages directly from an e-mail client and web pages from a web browser or Google Toolbar such as Text-to-voice which is an add-on to Firefox . Some specialized software can narrate RSS-feeds. On one hand, online RSS-narrators simplify information delivery by allowing users to listen to their favourite news sources and to convert them to podcasts. On the other hand, on-line RSS-readers are available on almost any PC connected to the Internet. Users can download generated audio files to portable devices, e.g. with a help of podcast receiver, and listen to them while walking, jogging or commuting to work.

A growing field in internet based TTS is web-based assistive technology, e.g. 'Browsealoud' from a UK company and Readspeaker. It can deliver TTS functionality to anyone (for reasons of accessibility, convenience, entertainment or information) with access to a web browser.

Other work is being done in the context of the W3C through the W3C Audio Incubator Group with the involvement of The BBC and Google Inc.

Others

- Some e-book readers, such as the Amazon Kindle.
- Some models of Texas Instruments home computers produced in 1979 and 1981 (Texas Instruments TI-99/4 and TI-99/4A) were capable of text-to-phoneme synthesis or reciting complete words and phrases (text-to-dictionary), using a very popular Speech Synthesizer peripheral. TI used a proprietary codec to embed complete spoken phrases into applications, primarily video games.
- IBM's OS/2 Warp 4 included VoiceType, a precursor to IBM ViaVoice.
- Systems that operate on free and open source software systems including Linux are various, and include open-source programs such as the Festival Speech Synthesis System which uses diphone-based synthesis (and can use a limited number of MBROLA voices), and gnuspeech which uses articulatory synthesis from the Free Software Foundation.
- Companies which developed speech synthesis systems but which are no longer in this business include BeST Speech (bought by L&H), Eloquent Technology

(bought by SpeechWorks), Lernout & Hauspie (bought by Nuance), SpeechWorks (bought by Nuance), Rhetorical Systems (bought by Nuance).
- GPS Navigation units produced by Garmin, Magellan, TomTom and others use speech synthesis for automobile navigation.

Speech synthesis markup languages

A number of markup languages have been established for the rendition of text as speech in an XML-compliant format. The most recent is Speech Synthesis Markup Language (SSML), which became a W3C recommendation in 2004. Older speech synthesis markup languages include Java Speech Markup Language (JSML) and SABLE. Although each of these was proposed as a standard, none of them has been widely adopted.

Speech synthesis markup languages are distinguished from dialogue markup languages. VoiceXML, for example, includes tags related to speech recognition, dialogue management and touchtone dialing, in addition to text-to-speech markup.

Applications

Speech synthesis has long been a vital assistive technology tool and its application in this area is significant and widespread. It allows environmental barriers to be removed for people with a wide range of disabilities. The longest application has been in the use of screen readers for people with visual impairment, but text-to-speech systems are now commonly used by people with dyslexia and other reading difficulties as well as by pre-literate children. They are also frequently employed to aid those with severe speech impairment usually through a dedicated voice output communication aid.

Software and sites such as CoolSpeech, TextSound, Ananova and YAKiToMe! have used speech synthesis to convert written news to audio content, which can be used for mobile applications.

Speech synthesis techniques are also used in entertainment productions such as games and animations. In 2007, Animo Limited announced the development of a software application package based on its speech synthesis software FineSpeech, explicitly geared towards customers in the entertainment industries, able to generate narration and lines of dialogue according to user specifications. The application reached maturity in 2008, when NEC Biglobe announced a web service that allows users to create phrases from the voices of Code Geass: Lelouch of the Rebellion R2 characters.

TTS applications such as CoolSpeech, TextSound, YAKiToMe! and Speakonia are often used to add synthetic voices to YouTube videos for comedic effect, as in Barney Bunch videos. TextSound and YAKiToMe! are also used to convert entire books for personal podcasting purposes, RSS feeds and web pages for news stories, and educational texts for enhanced learning.

Software such as Vocaloid can generate singing voices via lyrics and melody. This is also the aim of the Singing Computer project (which uses GNU LilyPond and Festival) to help blind people check their lyric input.

Next to these applications is the use of text to speech software also popular in Interactive Voice Response systems, often in combination with speech recognition.

Chapter 9

Voicemail

Voicemail (also known as **voice-mail**, **VMS**, or **message bank**) is a centralized system of stored telephone messages that can be retrieved later. The term is also used more broadly to denote any system of conveying a stored telecommunications voice message, including using an answering machine. Most cell phones have voicemail as a basic feature, and many land line phones and corporate PBXs have their own voicemail options.

Features

Voicemail systems are designed to convey a recorded audio message to a recipient. To do so they contain a user interface to select, play, and manage messages; a delivery method to either play or otherwise deliver the message; and a notification ability to inform the user of a waiting message. Most systems use phone-networks, either cellular or land-line based, as the conduit for all of these functions. Some systems may use multiple telecommunications methods, permitting recipients and callers to retrieve or leave messages through multiple methods.

Simple voicemail functions as a remote answering machine using a touch-tones as the user interface. More complicated systems may use other input devices such as voice or a computer interface. Simpler voicemail systems may play the audio message through the phone, while more advanced systems may have alternative delivery methods, including email or text message delivery, message transfer and forwarding options, and multiple mailboxes.

Notification methods also vary based on the voicemail system. Simple systems may not provide active notification at all, instead requiring the recipient to check with the system, while others may provide an indication that messages are waiting.

Almost all modern voicemail systems use digital storage and are typically stored on computer storage devices.

History

Voicemail was invented by Gerald M. Kolodny and Paul Hughes, where it was first described in an article in the medical journal, Radiology (Kolodny GM, Cohen HI, Kalisky A. Rapid-access system for radiology reports: a new concept. Radiology. 1974;111(3):717-9) A patent was applied for by Kolodny and Hughes in 1975, and was issued in 1981 (U.S. patent 4,260,854). The patent was assigned to Sudbury Systems of Sudbury Massachusetts who proceeded to market and sell such systems to corporations and hospitals. IBM, Sony and Lanier, as well as several smaller makers of voicemail systems, licensed the Sudbury patent for their voicemail systems. A patent suit, brought by Pitney Bowes, claiming prior art to the Sudbury patent, was denied by the U.S. District Court, District of Connecticut on November 8, 2000. In the 1970s and early 1980s, the cost of making a phone call decreased and more business communication was done by phone. As corporations grew and labor rates increased, the ratio of secretaries to employees decreased. With multiple time zones, fewer secretaries and more communication by phone, real-time phone communications were hampered by callers being unable to reach people. Some early studies showed that only 1 in 4 phone calls resulted in a completed call and half the calls were one-way in nature (that is, they did not require a conversation). This happened because people were either not at work (due to time zone differences, being away on business, etc.), or if they were at work, they were on the phone, away from their desks in meetings, on breaks, etc. This bottleneck hindered the effectiveness of business activities and decreased both individual and group productivity. It also wasted the caller's time and created delays in resolving time-critical issues.

Neither email messaging nor cellular phones were widespread in the 1970s and 80s, and did not really begin to flourish until the mid-1990s. The initial solution to the phone communication problem for businesses was the "message center." A message center or "message desk" was a centralized, manual answering service inside a company staffed by a few people answering everyone's phones. Extensions that were busy or rang "no answer" would forward to the message center onto a device called a "call director". The call director had a button for each extension in the company which would flash when that person's extension forwarded to the message center. A little label next to the button told the operator whose extension it was.

While it was an improvement over earlier systems, the message center had many disadvantages. Operators were busy, and volumes of calls would come in simultaneously at peak periods, such as lunch time. This left message attendants with little time to take each message accurately. Often, they also weren't familiar with employees' names or how to spell or pronounce them. Messages were written on pink slips and distributed by the internal mail system. The messages often arrived at people's desks after lengthy delays, contained little content other than the caller's name and number, and were often inaccurate, with misspelled names and wrong phone numbers.

Tape-based telephone answering machines had come into the residential telephone market, but they weren't used much in the corporate environment due to physical

limitations of the technology. One answering machine was needed for each telephone; messages couldn't be recorded if the user was using the phone; messages had to be retrieved in sequential order; and messages couldn't be retrieved remotely, selectively discarded, saved, or forwarded to others. Further, the manufacturers of PBXs (private branch exchanges — the name for corporate phone systems) used proprietary digital phone sets in order to increase the functionality and value of the PBX. These phone sets were, by design, incompatible with answering machines.

Corporate Voicemail was broadly commercialized by Octel Communications (founded in 1982 by Bob Cohn and Peter Olson). ROLM Corporation (founded in 1969 by Gene Richeson, Ken Oshman, Walter Loewenstern and Robert Maxfield and later owned by IBM before IBM sold it to Siemens) was the first PBX manufacturer to offer integrated voicemail with its PhoneMail system, and also played a major role commercializing voicemail.

IBM's product, initially called the SFS (Speech Filing System) was developed as an intensive research project at the IBM Thomas J. Watson Research Center. It was meant to mimic the concept of email, but using the telephone as the input device and the human voice as the medium for the message. Work on the system began in 1973 and the first operational prototype was made available to users in 1975. Four people could use it at once. From 1975-1981, about 750 IBM executives, mainly in the U.S., used various SFS prototypes in their daily work. Those prototypes ran on an IBM System /7 computer attached to an IBM VM370 for additional storage. The prototype was converted to run on a Series /1 computer in 1978. In September, 1981, IBM announced this product as the "Audio Distribution System" (ADS) with the first customer installation being February, 1982. It was marketed directly by IBM and for a short while by AT&T. IBM's ADS required special attention as a computer (special room, special power, air conditioning, etc.) ADS was richly featured for voice messaging, the result of IBM's enormous research in human factors and observing SFS in real operational use. However, ADS had major limitations which resulted in its failure as a commercial product: for example, it was physically large, expensive, limited to 1,000 users, had no telephone answering mode (could not answer outside calls), and had to be taken out of service to make administrative changes to the user data base (called "MAC", for "moves, adds and changes").

Another company, Delphi Communications of California, deserves some partial credit for invention of voicemail. Under the leadership of Jay Stoffer, Delphi developed a proprietary system (called Delta 1) that picked up incoming calls directly from the telephone company. Stoffer presented the Delphi concept publicly to the association of Telephone Answering Services around 1973 and the prototype system was launched in San Francisco in 1976 by a Delphi company called VoiceBank. Delphi developed Delta 1 as a purely service-oriented voice messaging system to answer subscriber telephones for businesses and professionals. Delta 1 required human intervention for message deposit. While three machines were built, only one machine was put into operational service. The completely automated voice messaging system (Delta 2) was developed for initial operational use in Los Angeles in 1981. Apparently Delta 2 was built, installed and

operational for a short while, but unfortunately Delphi's major early investor, Exxon Enterprises, abruptly shut down Delphi in July, 1982. Nothing further was done with Delphi's technology. A patent was applied for and issued for Delphi's Automated Telephone Voice Service System. The patent, U.S. Patent No. 4,625,081, was issued after Delphi's closure, but Delphi's assets (and the patent) were transferred to another Exxon company, Gilbarco, which made equipment for gas pumps at filling stations. Gilbarco is now owned by GEC in the United Kingdom.

In 1979, five years after IBM's SFS (ADS) system and three years after Delphi's Delta 1 system were first operational, a company was founded in Texas by Gordon Matthews called ECS Communications (the name was later changed to VMX). According to Jay Stoffer, founder of Delphi Communications, Gordon Matthews learned about Delphi's voicemail prior to his founding VMX. Regardless of how he was inspired, Matthews eventually founded VMX which developed a 3,000-user voice messaging system called the VMX/64. VMX was arguably the first company to offer voicemail for sale commercially for corporate use. Matthews was able to sell his system to several notable large corporations, such as 3M, Kodak, American Express, Intel, Hoffmann–La Roche, Corning Glass, Arco, Shell Canada and Westinghouse. This impressive list of early adopters started the ball rolling on corporate voicemail. While some claim that VMX and Gordon Matthews invented voicemail, this claim is not true. The first inventor of record was Stephen Boies of IBM years before VMX was founded.

While VMX began with a good start, it failed at developing the market, and the company was not a commercial success. It took many years before its products could answer outside calls (and then only under certain circumstances), they were physically enormous, expensive, light on important user features and had serious reliability issues. In addition, the user interface was cumbersome, requiring the users to remember non-intuitive multi-digit Touch-tone commands. Matthews, a prolific entrepreneur and patenter, applied for and was granted a patent on voicemail (patent number 4,371,752) which issued in February, 1983. The patent was promoted as the pioneering patent for voicemail.

Shortly after the development of the first voicemail systems, several companies sprang up to develop their own systems including Wang Laboratories, ROLM, Opcom, Octel, Centigram, Genesis, and many others. Wang Laboratories, under the leadership of Dr. Larry Bergeron, developed a voicemail system modeled after the IBM system. Wang called its system the DVX. It too could not answer outside calls but was smaller and less expensive than the IBM system.

Matthews was quite astute at the way he used his patent. Matthews tried to assert his patent with IBM, AT&T and then Wang, but all three companies reportedly would have been able to invalidate the Matthews patent because of prior art. Matthews cleverly achieved a settlement where the patent was let stand, not challenged in court and IBM, Wang and AT&T (in separate settlements) received royalty-free licenses to all VMX patents. Wang, the last of the majors to get such a license, essentially paid $20,000 and cross licensed a few patent applications (not issued patents). IBM and AT&T also cross-licensed a number of patents to VMX, most of which were obsolete or outdated. VMX

could claim that several major companies licensed the patent (even though they paid almost nothing to VMX for the rights), but that part wasn't disclosed. The patent was never challenged in court and VMX then continued to assert (incorrectly) that it had invented voicemail and that Matthews was the father of voicemail. Following the settlement with Wang, VMX settled with Octel. In exchange for a small payment and Octel's agreeing not to litigate any VMX patent, Octel received a paid-up, royalty-free license on all existing and future VMX patents.

ROLM (one of the first makers of digital PBX's) was the first company to offer integrated voicemail through its product called PhoneMail, which name is a registered trademark. PhoneMail offered impressive recording quality of its digitized messages. ROLM's digital PBX (called a CBX, for Computerized Branch eXchange) was the first PBX to provide signaling to indicate which extension was being forwarded to a voicemail system (the first PBX to do so). However, the signaling was proprietary and intended only for use by its voicemail product, PhoneMail. ROLM's CBX also provided signaling to enable PhoneMail to illuminate a message waiting light on ROLM's electronic phones and later standard phones equipped with message waiting lights (also a studder dialtone is used with analog and digital phones). PhoneMail worked with most but not all models of ROLM's CBXs, would work with some other brands of PBXs such as Nortel's Option Meridian (with adaptors and loss of some features), and was heavily promoted by ROLM. PhoneMail is still a commercial success. Siemens still offers PhoneMail in various configurations/sizes (including a micro-sized version) and its unified messaging successor, Xpressions 470; along with the same pleasing female voice most ROLM techs have nicknamed, Silicon Sally.

Opcom, a company started by David Ladd, was another maker of voicemail which also pioneered and patented the feature of automated attendant(U.S. Patent numbers 4,747,124 and 4,783,796 both of which issued in 1988). Automated attendant is not technically voicemail, but all the features to enable automated attendant are already part of a voicemail system so it is a natural feature to add to it. Despite the Opcom patents, automated attendant features had been implemented in a commercial Teradyne product prior to 1985.

Opcom developed a voicemail system primarily marketed to smaller enterprises. Automated attendant enables callers to direct calls by pressing single digit keys. For example, "If you are making domestic reservations, press 1; for international reservations, press ‘2'; for frequent flier information, press ‘3', etc." Opcom was an innovative company and also pioneered the concept of Unified Messaging. Opcom's voicemail product was a commercial success with smaller companies and some large ones. Around 1991, VMX was on the verge of bankruptcy and was acquired by Opcom. Since Opcom was private and VMX was public, the transaction was done as a reverse merger and the surviving company was called VMX. Little of the original VMX Company was retained. Within a few years, VMX was acquired by Octel and David Ladd became Octel's Chief Technology Officer.

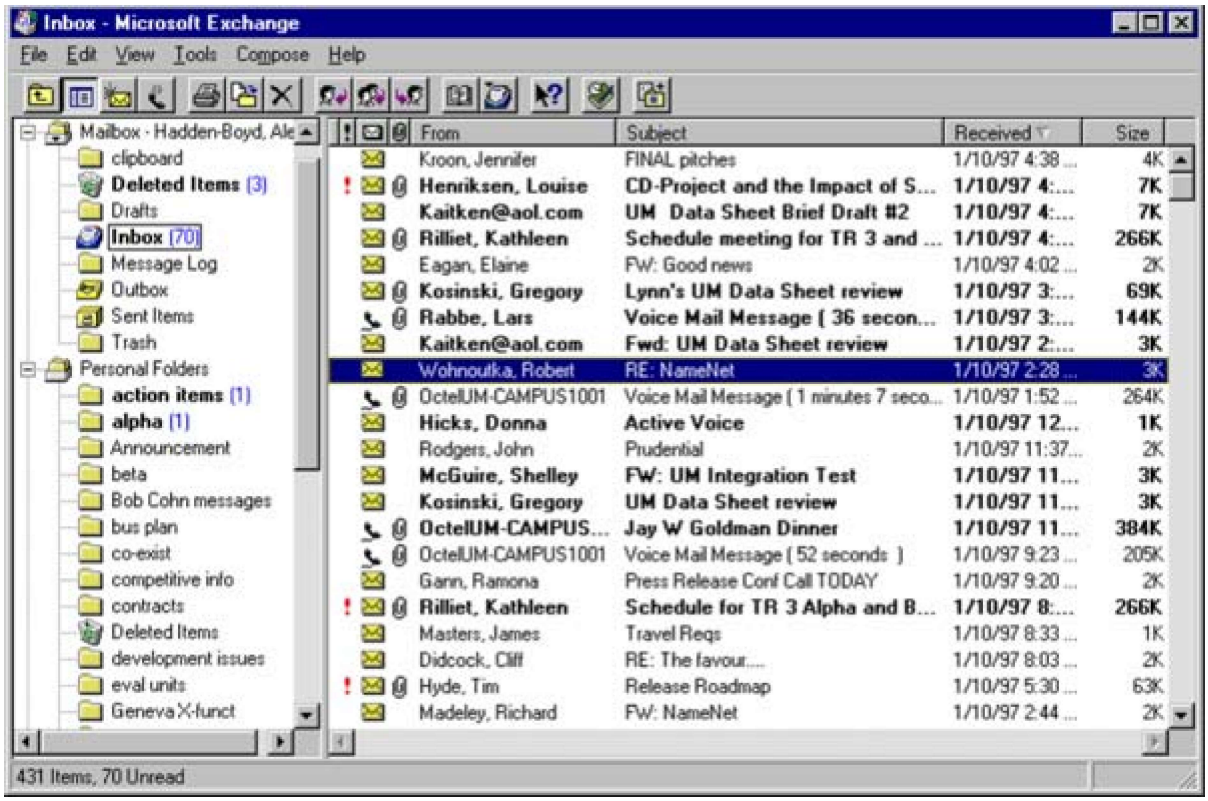
Octel Communications Corporation was founded in Silicon Valley in 1982 by Bob Cohn and Peter Olson. Octel's voicemail system (developed during the period from 1982–1984 and first sold in 1984), became the clear market leader fairly quickly. While Octel benefited from the work and experiments of others, it was the first stand-alone voicemail company to build a strong business and strategy to win at this important market. In addition, Octel innovated substantially new technology which contributed heavily to its success. Octel's differentiated hardware and software architecture enabled its systems to be physically smaller, faster, more reliable, and much less costly to build than any other vendor. These features, many of which were patented, gave Octel market leadership:

- User-friendly user-interface (other systems were not intuitive and had no help prompts).
- Error-free Touch-tone detection (other systems falsely detected a Touch- tone out of human voice, or didn't detect Touch-tones when users pressed the buttons).
- Scrambled messages so no one could hear anyone else's messages (other systems could accidentally get other people's messages if the system failed at the right time).
- Telephone answering, voice messaging and automated attendant.
- Moves, adds and changes could be done while the system was running.
- Large amounts of message storage.
- Physically small size (about the size of a 2-drawer filing cabinet, compared to ROLM's original PhoneMail being about 5' × 5' × 5' and VMX's system filling a computer room). No requirement for special environment.
- Locatable anywhere. Octel systems could be located in any office environment and they were not susceptible to electrical shocks (often common on carpeted floors in offices, especially during winter).
- High reliability (being the first voicemail system to achieve up-time of 99.9% with its first system).
- Compatible with virtually all brands of PBX (voicemail offered by PBX vendors could only work with that vendor's PBX system).
- Telephone answering with all PBXs, even those which had no method of providing caller ID.
- Message notification (phoning subscribers at various locations pre-programmed by the subscriber, when messages were received).
- Range of capacities. Small, medium, large and extra large capacity systems that addressed the needs of major companies (For example, Octel's systems had 50% greater port capacity than VMX's largest system). Small systems went in branch offices, medium systems went in district offices, large systems went in regional offices, and extra large systems could handle large corporate headquarters with over 10,000 people.
- Networking between voicemail systems so companies could have their voicemail systems operate as one large virtual network.

Octel's strategy addressed needs of major accounts which other vendors did not until much later: advanced training, customer service, sales and market education. Octel's system could identify the extension number of calls being forwarded to it and light

message-waiting lights on most PBXs. This was possible because Octel's engineers reverse engineered the major brands of PBX (legally) and often figured out ways to communicate with the PBX in ways the PBX manufacturer had not. Eventually most makers of PBX chose to work cooperatively with Octel. Octel integrated with almost 100 brands of PBX worldwide. As a result of Octel's worldwide leadership, its user interface (which was done in more than 75 languages and dialects) became the most widely known in the world.

Toward the late 1990s, Octel introduced the concept of Visual Mailbox and *Unified Messaging*. Visual Mailbox enabled users to manage their voice mailboxes through their PCs, although the messages were still stored on the Octel system. Unified Messaging integrated voicemail into Microsoft Exchange, the corporate email system made by Microsoft. Unified Messaging had actually been invented by Roberta Cohen, Kenneth Huber and Deborah Mill at AT&T Bell Labs. The patent for Unified Messaging was received in June, 1989 (Patent number 4,837,798).



The figure above shows a screen shot from an early Unified Messaging system (GUI). Emails are identified with the icon of an envelope; voicemails are identified with the icon of a phone handset. This system fully integrated voicemail into Microsoft Exchange so both voicemails and emails could be displayed and accessed via Microsoft Outlook. Users could also call into the system by phone and hear both voicemails and machine-read emails (TUI).

Unified Messaging: With Unified Messaging, users could access voice and email messages using either the graphical user interface (GUI) on their PC, or using the telephone user interface (TUI) with any telephone in the world. On the PC, users could see voicemails and emails mixed together in their email inbox. Voice mails had a little telephone icon next to them and emails had a little envelope icon next to them. For voicemail, they'd see the "header information" (sender, date sent, size, and subject). Users could double-click a voicemail from their email inbox and hear the message through their PC or a phone next to their desk. Using any phone in the world, users could listen to voice messages like they normally did, plus have emails read to them (in synthesized voice). Voice messages could be sent using email or telephone addressing schemes, and the data networking infrastructure was used to send messages between locations rather than the public switched telephone network. Unified Messaging was not a commercial success at the time because in the late 1990s email did not enjoy a huge market share, email servers were not very reliable, internet connections were slow (voice messages were large files) and most PCs did not have speakers or microphones.

Until 1988, telephone companies and the newly formed cellular phone companies were barred by law from offering voicemail to their subscribers. This was done by the FCC to protect the telephone answering businesses around the country. This prohibition continued with the decree which broke up AT&T in 1984. A subsequent ruling by Judge Harold H. Greene on March 7, 1988, reversed this barrier. Phone companies were allowed to offer voicemail as a service, but they were barred from designing or manufacturing the machines that could provide the service.

VMX's large system was used by a few carriers (telephone companies), but severe reliability and cost issues prevented VMX from expansion to the carrier market. Octel already had very high capacity systems for corporate use and by 1988 all seven Regional Bell Operating companies were using Octel for internal use. Octel first adapted its largest system for the carriers, which enabled them to offer reliable voicemail to their subscribers. Within a year, Octel launched a new generation of its large system specifically designed for carriers which was compliant with "NEBS standards," the tight standard required by phone companies for any equipment located in their central offices. A few other manufacturers entered the voicemail market for carriers including Unisys, Boston Technology, and Comverse Technology (an Israeli based company founded by Kobi Alexander). These vendors did not offer voicemail to corporations but they focused on the potentially large and lucrative carrier market. Unisys secured PacBell's residential voicemail services, and Boston Technology was the mainstay of Bell Atlantic's residential voicemail offering. None of the other corporate voicemail manufacturers had notable success with the carrier market because their systems' capacities were too small and the equipment wasn't reliable enough. Selling to carriers also required a different method of sales and marketing than selling to the corporate market, and only Octel succeeded at both.

Perhaps the first cellular carrier in North America to offer voicemail successfully to its subscribers was Bell Cellular, the Canadian carrier serving Ontario and Quebec (Bell Cellular later changed its name to Bell Mobility). Bell Cellular's success with voicemail

caught on, and cellular voicemail spread throughout Canada and then to the US and overseas. Within a few years, 100% of Canadian cellular companies ultimately used Octel voicemail, followed by virtually all of the major US wireless carriers (including the seven RBOCs, AT&T Wireless and McCaw) and a large percent of the GSM carriers around the world. Comverse Technology was very successful in the GSM market outside the US. The Octel user interface became the most common in the world with carriers, but each carrier made minor variations on the interface.

Other interesting markets developed from the carrier market including a concept called "*virtual telephony*." Virtual Telephony, developed by Octel, used voicemail to provide phone service rapidly in emerging countries without wiring for telephones. The problem this solved was that emerging countries did not have many telephones. Wiring for telephones was very expensive, and many poorer citizens didn't have homes to wire. The economies of emerging countries were held back partly because people couldn't communicate beyond the area where they could walk or ride a bicycle. Giving them phones was one way to help their economies, but there wasn't a practical way to do it. In some countries, the wait for a phone was several years and the cost was in the thousands of dollars. Cellular phones weren't an option at the time because they were extremely expensive (thousands of dollars per handset) and the infrastructure to install cell sites was also costly.

With virtual telephony, each person could be given a phone number (just the number, not the phone) and a voice mailbox. The citizen would also be given a pager. If someone called the phone number, it never rang on an actual phone, but would be routed immediately to a central voicemail system. The voicemail system answered the call and the caller could leave a long, detailed message. As soon as the message was received, the voicemail system would trigger the citizen's pager. When the page was received, the citizen would find a pay phone and call in to pick up the message. This concept was used successfully in South America and South Africa.

1980s-1990s

In the early 1980s there were over 30 companies vying for the corporate voicemail market including many companies no longer in business today. Among the many contenders were IBM, VMX, Wang, Octel, ROLM, AT&T, Northern Telecom, Delphi Communications, Voice and Data Systems, Opcom, Commterm, Genesis, Brook Trout, Glenayre, BBL, AVT, AVST, Digital Sound, Centigram, Voicemail International, Active Voice, and many others. Virtually all contenders in the corporate voicemail market were based in the United States.

By the mid 1990s, IBM and Wang exited the voicemail market because they couldn't get enough traction. ROLM was purchased by IBM in the mid 1980s (which was a financial disaster for the profitable ROLM, as IBM clearly could not grasp the laid back, "think outside the box" attitude of ROLM, which was the #2 PBX supplier in the US from the mid 70s to late 80s), then sold half interest to the German company, Siemens. In 1992, Siemens bought ROLM entirely from IBM and the original ROLM product line was done

for, except for PhoneMail (the only product Siemens did not destroy). VMX suffered from poor product and ineffective management and was about to fold when Opcom merged with it. The surviving company was called VMX, but VMX was all but erased by Opcom except for its name and patent portfolio. In 1994, Octel bought VMX. By the early 90s, AT&T/Lucent created its version of voicemail for the corporate market (called Audix) but it would only work on AT&T/Lucent PBXs. Nortel developed Meridian Mail and followed the same strategy as AT&T in that Meridian Mail only worked with Northern Telecom PBXs. As a result, neither company achieved much market share with large national or multi-national accounts (because few major companies, if any, used only one brand of PBX, though Nortel had been the major leader since the late 1970s with ROLM a close second and poised to overtake Nortel until IBM bought ROLM). AT&T spun off its equipment business into a company called Lucent Technologies, and Northern Telecom changed its name to Nortel. Several small companies offering voicemail folded because of inadequate product or management.

By the mid-1990s, Octel had become the number one supplier of voicemail both to corporations and to carriers. It had about a 60% market share in the U.S., Canada, Europe and Japan (for large corporations) and between a 30% and 100% of the carrier market, depending on the country. By 1997 Octel's biggest competitors were Audix, made by Lucent, and Meridian Mail, made by Nortel. In July 1997, Octel was purchased by Lucent Technology. Lucent's AUDIX division was merged into Octel to form the Octel Messaging Division. In the same year, Boston Technology was acquired by Comverse Technology making it the second largest supplier to carriers after Octel. In a few years Comverse became the largest supplier to carriers with Lucent holding its leadership in the corporate market and second place with carriers. By 2000, some estimate that there were over 150,000,000 active users of corporate and carrier voicemail made by the Octel Messaging Division. Shortly thereafter, Lucent spun off its corporate business, including the Octel Messaging Division, into a company known as Avaya. Comverse today retains its leadership of legacy voicemail systems sold to carriers around the world. For IP based voicemail systems Ericsson claims market leadership with its Ericsson Messaging-over-IP (MoIP) solution.

2000s

By the year 2000, voicemail had become a ubiquitous feature on phone systems serving companies, cellular and residential subscribers. Cellular and residential voicemail continue today in their previous form, primarily simple telephone answering. Email became the prevalent messaging system, email servers and software became quite reliable, and virtually all office workers were equipped with multimedia desktop PCs.

Instant messaging in voice: The next development in messaging was in making text messaging real-time, rather than just asynchronous store-and-forward delivery into a mailbox. It started with Internet service provider America Online (AOL) as a public Internet-based free text "chat" service for consumers, but soon was being used by business people as well. It introduced the concept of Internet Protocol "presence management" or being able to detect device connectivity to the Internet and contact

recipient "availability" status to exchange real-time messages, as well as personalized "Buddy list" directories to allow only people you knew to find out your status and initiate a real-time text messaging exchange with you. Presence and Instant Messaging has since evolved into more than short text messages, but now can include the exchange of data files (documents, pictures) and the escalation of the contact into a voice conversational connection.

Mobile devices

The increase in wireless mobility, originally through cellular services and today through IP-based Wi-Fi, was also a driver for messaging convergence with mobile telephony. Today it is not only fostering the use of speech user interfaces for message management, but increasing the demand for retrieval of voice messages integrated with email. It also enables people to reply to both voice and email messages in voice rather than text. New services, such as GotVoice, SpinVox and YouMail, are helping to blur the boundaries between voicemail and text by delivering voicemails to mobile phones as SMS text messages.

Unified messaging with VoIP

Corporate voicemail did not change much until the advent of Voice over IP (VoIP — voice being transmitted over the internet) and the development of IP telephony applications to replace legacy PBX telephony (called TDM technologies). IP (Internet Protocol) telephony changed the style and technology of PBXs and the way voicemail systems integrated with them. This, in turn, facilitated a new generation of Unified Messaging, which is now likely to catch on widely. The flexibility, manageability, lower costs, reliability, speed, and user convenience for messaging convergence is now possible where it wasn't before. This might include intra- and inter-enterprise contacts, mobile contacts, proactive application information delivery, and customer contact applications.

The corporate IP telephony-based voicemail CPEmarket is served by several vendors including Avaya, Cisco systems, Adomo, Interactive Intelligence, Nortel, Mitel, 3Com, and AVST. Their marketing strategy will have to address the need to support a variety of legacy PBXs as well as new Voice over IP as enterprises migrate towards converging IP-based telecommunications. A similar situation exists for the carrier market for voicemail servers, currently dominated by Comverse Technology, with some share still held by Lucent Technologies.

VoIP telephony enables centralized, shared servers, with remote administration and usage management for corporate (enterprise) customers. In the past, carriers lost this business because it was far too expensive and inflexible to have remote managed facilities by the phone company. With VoIP, remote administration is far more economical. This technology has re-opened opportunities for carriers to offer hosted, shared services for all forms of converged IP telecommunications, including IP-PBX and voicemail services. Because of the convergence of wired and wireless communications, such services may also include support of a variety of multi-modal handheld and desktop end user devices.

This service, when offered for multiple extensions or phone numbers is sometimes also called Unified Voicemail.

Business efficiency

Voice mail's introduction enabled people to leave lengthy, secure and detailed messages in natural voice, working hand-in-hand with corporate phone systems. The adoption of voicemail in corporations improved the flow of communications and saved huge amounts of money. GE, one of the pioneer adopters of voicemail in all of its offices around the world, claimed that voicemail saved, on average, over US$1,100 per year per employee.

Voicemail has two main modes of operation: *telephone answering* and *voice messaging*. Telephone answering mode answers outside calls and takes a message from any outside caller (either because the extension was busy or rang no-answer). Voice messaging enables any subscriber (someone with a mailbox number) to send messages directly to any or many subscribers' mailboxes without first calling them. Both of these modes are described below.

Telephone answering mode

One of the advantages of a PBX is its ability to forward calls. If a person is using his phone or does not answer it, calls to his extension are forwarded automatically by the PBX to another extension, presumably someone (like a secretary) who can answer the call and take a message. With a voicemail system installed, the PBX is programmed to forward busy or unanswered extensions to a machine — the voicemail system.

Suppose an outside caller, Willma, calls someone in a company, Fred. If Fred's phone rings "no answer" or "busy", the PBX will forward the call to the voicemail system. Somehow the PBX needs to tell the voicemail system that Fred's phone is the one that the call is being forwarded to so that the voicemail system can answer with Fred's personal greeting. Without this information, the voicemail system would have no idea whose phone it was answering. Once a message is left, the voicemail system illuminates the message waiting light on Fred's phone. It does this by sending a signal to the PBX to tell it which light to light. When Fred returns to his desk and calls the voicemail system (or calls in remotely) he is presented only with the messages in his personal mailbox even though thousands of messages belonging to other people are stored on the same system. Once the messages are played, the voicemail system signals the PBX to turn off the message waiting light on Fred's phone.

Early voicemail systems (notably those made by IBM and VMX) could not answer outside calls — that is, they could not automatically answer a call originally destined to an extension on the PBX which rang busy or was not answered. As subsequent voicemail systems emerged (notably ROLM and Octel), the systems could answer outside calls. However, most PBX's did not provide signaling to tell the voicemail system which extension it was forwarding, nor did they support telephones with message waiting lights.

This signaling would come later, but until it did it created a major challenge for voicemail systems for many years.

Interoperability between systems

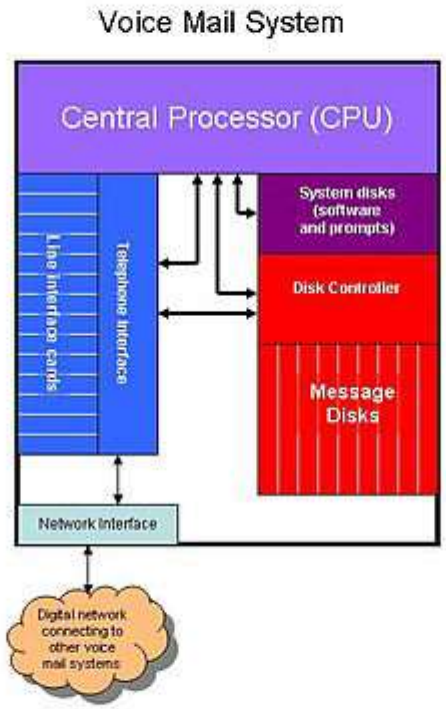
Voice messaging does not always have to be sent between individuals on the same voicemail system. Messages can be transferred using AMIS (Audio Messaging Interchange Specification) or VOIP (Voice Over Internet Protocol) technologies; both allow messages on one computer system to be forwarded to the target system. Like email, this method of delivering voice messages can be subject to abuse such as spam or vishing. There are Federal and State laws and regulations designed curb these abuses, such as the United States National Do Not Call Registry.

How voicemail systems work

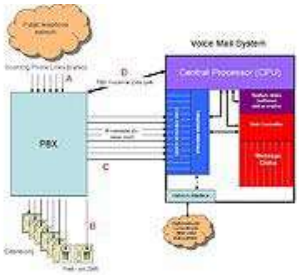
Here we, describes how the original style, standalone, voicemail system worked with a corporate PBX. The principle is the same with Central Office Switches (CO Switches) or Mobile Telephone Switching Office (MTSOs). More modern voicemail systems work on the same principle, but some of the components may be shared with other systems, such as email systems.

Voicemail systems contain several elements shown in the figure below:

- A central processor (CPU) which runs the operating system and a program (software) that gives the system the look-and-feel of a voicemail system. This software includes thousands of pre-recorded prompts that "speak" to the users as they interact with the system;
- Disk controller and multiple disk drives for message storage;
- System disks which not only include the software above, but also contain a complete directory of all users with pertinent data about each (name, extension number, voicemail preferences, and pointers to each of the messages stored on the message disk that belong to them);
- Telephone interface system that enables many phone lines to be connected to it.



The drawing below shows how the voicemail system interacts with the PBX. Suppose an outside caller is calling Fred's extension 2345. The incoming call comes in from the public network (A) and comes into the PBX. The call is routed to Fred's extension (B), but Fred doesn't answer. After a certain number of rings, the PBX stops ringing Fred's extension and forwards the call to an extension connected to the voicemail system (C). It does this because PBXs are generally programmed to forward busy or unanswered calls to another extension. Simultaneously the PBX tells the voicemail system (through signaling link D) that the call it is forwarding to voicemail is for Fred at extension 2345. In this way, the voicemail system can answer the call with Fred's greeting.



There are many microprocessors throughout the system since the system must handle large amounts of data and it's unacceptable to have any wait times (for example, when the

system is recording or playing your message, it's unacceptable if the system stops recording momentarily like computers often do while accessing large files).

When Fred's extension forwards to the voicemail system, the Telephone Interface detects ringing. It signals to the Central Processor (CPU) that a call is coming in. The CPU simultaneously receives a signal on the PBX-Voicemail Data Link (D) telling it that extension 2345 is being forwarded on ring-no-answer to the specific extension that is now ringing. The CPU directs the Telephone Interface (which controls the line interface cards) to answer the call. The CPU's program realizes that it's a call for Fred so it looks up Fred's greeting immediately and directs the Disk Controller to start playing it to the caller. It also plays some system prompts instructing the caller what comes next (for example, "When you have finished recording, you may hang up or press '#' for more options"). All "talking" to the caller is done through prompts that are selected by the CPU according to the program stored in the voicemail system. The CPU selects the prompts in response to the keys the caller presses.

The caller's message is digitized by the Telephone Interface system and transmitted to the Disk Controller for storage onto the Message Disks. Some voicemail systems will scramble the message for further security. The CPU then stores the location of that message in the System Disk inside Fred's mailbox directory entry. After the caller hangs up and the message has been stored, the CPU sends a signal to the PBX through the link (D) instructing the PBX to turn on the message waiting light on Fred's phone.

When Fred comes back to his desk and sees the light on his phone, he calls a designated extension number for the voicemail system (an actual extension number assigned to the lines in "C" in the figure above).

Again the Telephone Interface alerts the CPU that a call is coming in on a particular line, but this time the signaling from the PBX-Voicemail Data Link (D) indicates that Fred is calling directly, not being forwarded. The CPU directs the Telephone Interface to answer the call.

Since the CPU "knows" it is Fred (from the signaling on the Data Link D), it looks up Fred's information on the System Disk, specifically his password. The CPU then directs Disk Controller to play a log-on prompt to the user: "Please enter your password." Once the password is entered (via Touch-tones), the CPU compares it to the correct one and, if entered correctly, allows Fred to continue.

The CPU then determines (from Fred's directory entry) that Fred has a new message. The CPU then presents Fred his options (e.g., "You have a new message. To listen to your new message, press 1; to record a message, press 2" etc.) The options are presented by the CPU directing the Disk Controller to play prompts, and the CPU listens for Touch-tones from Fred. This interaction of playing prompts and responding with Touch-tones enables Fred to interact with the voicemail system easily.

If Fred presses 1 to listen to his message, the CPU looks up the location of Fred's new message in his mailbox directory (on the System Disk), and directs the Disk Controller to play that message. The Disk Controller finds the message on the Message Disks, and sends the data stream directly to the Telephone Interface. The Telephone Interface then converts the data stream to sound and plays it to Fred through the Line Interface Card which Fred is connected to.

Playback controls (like rewind, pause, fast forward, changing volume, etc.) are all input via Touch-tones, are "read" by the CPU, and the appropriate actions are taken based on the stored program in the system. For example, if Fred wants to pause message playback, he might press 2. Since the CPU is constantly listening for Touch-tones from Fred, his command causes the CPU to direct the Disk Controller to stop playing the message. A variety of playback controls and options are available on most sophisticated voicemail systems so that users can control message playback, store messages in archives, send messages to groups, change their preferences, etc.

The better designed voicemail systems have a user-friendly interface with clear and meaningful prompts so the interaction with the voicemail system is quick and easy.

Chapter 10

Voice Over IP

Voice over Internet Protocol (**Voice over IP**, **VoIP**) is one of a family of internet technologies, communication protocols, and transmission technologies for delivery of voice communications and multimedia sessions over Internet Protocol (IP) networks, such as the Internet. Other terms frequently encountered and often used synonymously with VoIP are *IP telephony*, *Internet telephony*, *voice over broadband* (VoBB), *broadband telephony*, and *broadband phone*.

Internet telephony refers to communications services—voice, fax, SMS, and/or voice-messaging applications—that are transported via the Internet, rather than the public switched telephone network (PSTN). The steps involved in originating a VoIP telephone call are signaling and media channel setup, digitization of the analog voice signal, encoding, packetization, and transmission as Internet Protocol (IP) packets over a packet-switched network. On the receiving side, similar steps (usually in the reverse order) such as reception of the IP packets, decoding of the packets and digital-to-analog conversion reproduce the original voice stream.

VoIP systems employ session control protocols to control the set-up and tear-down of calls as well as audio codecs which encode speech allowing transmission over an IP network as digital audio via an audio stream. The codec used is varied between different implementations of VoIP (and often a range of codecs are used); some implementations rely on narrowband and compressed speech, while others support high fidelity stereo codecs.

Protocols

Voice over IP has been implemented in various ways using both proprietary and open protocols and standards. Examples of technologies used to implement Voice over IP include:

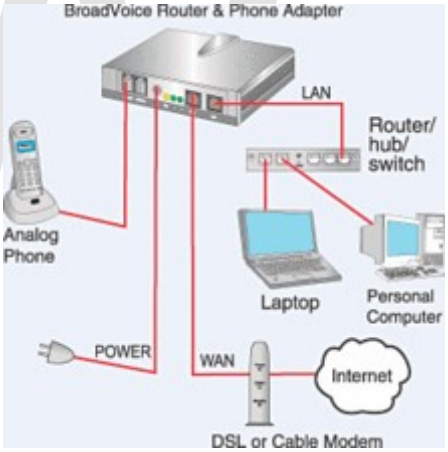
- H.323
- IP Multimedia Subsystem (IMS)
- Media Gateway Control Protocol (MGCP)
- Session Initiation Protocol (SIP)
- Real-time Transport Protocol (RTP)
- Session Description Protocol (SDP)

The H.323 protocol was one of the first VoIP protocols that found widespread implementation for long-distance traffic, as well as local area network services. However, since the development of newer, less complex protocols, such as MGCP and SIP, H.323 deployments are increasingly limited to carrying existing long-haul network traffic. In particular, the Session Initiation Protocol (SIP) has gained widespread VoIP market penetration.

A notable proprietary implementation is the Skype protocol, which is in part based on the principles of Peer-to-Peer (P2P) networking.

Adoption

Consumer market



Example of residential network including VoIP

A major development that started in 2004 was the introduction of mass-market VoIP services that utilize existing broadband Internet access, by which subscribers place and receive telephone calls in much the same manner as they would via the public switched telephone network (PSTN). Full-service VoIP phone companies provide inbound and outbound service with Direct Inbound Dialing. Many offer unlimited domestic calling for a flat monthly subscription fee. This sometimes includes international calls to certain countries. Phone calls between subscribers of the same provider are usually free when flat-fee service is not available.

A VoIP phone is necessary to connect to a VoIP service provider. This can be implemented in several ways:

- Dedicated VoIP phones connect directly to the IP network using technologies such as wired Ethernet or wireless Wi-Fi. They are typically designed in the style of traditional digital business telephones.
- An analog telephone adapter is a device that connects to the network and implements the electronics and firmware to operate a conventional analog telephone attached through a modular phone jack. Some residential Internet gateways and cablemodems have this function built in.
- A softphone is application software installed on a networked computer that is equipped with a microphone and speaker, or headset. The application typically presents a dial pad and display field to the user to operate the application by mouse clicks or keyboard input.

PSTN and mobile network providers

It is becoming increasingly common for telecommunications providers to use VoIP telephony over dedicated and public IP networks to connect switching stations and to interconnect with other telephony network providers; this is often referred to as "IP backhaul".

Smartphones and Wi-Fi enabled mobile phones may have SIP clients built into the firmware or available as an application download. Such clients operate independently of the mobile telephone phone network and use either the cellular data connection or WiFi to make and receive phone calls.

Corporate use

Because of the bandwidth efficiency and low costs that VoIP technology can provide, businesses are gradually beginning to migrate from traditional copper-wire telephone systems to VoIP systems to reduce their monthly phone costs.

VoIP solutions aimed at businesses have evolved into "unified communications" services that treat all communications—phone calls, faxes, voice mail, e-mail, Web conferences and more—as discrete units that can all be delivered via any means and to any handset, including cellphones. Two kinds of competitors are competing in this space: one set is focused on VoIP for medium to large enterprises, while another is targeting the small-to-medium business (SMB) market.

VoIP allows both voice and data communications to be run over a single network, which can significantly reduce infrastructure costs.

The prices of extensions on VoIP are lower than for PBX and key systems. VoIP switches may run on commodity hardware, such as PCs or Linux systems. Rather than closed architectures, these devices rely on standard interfaces.

VoIP devices have simple, intuitive user interfaces, so users can often make simple system configuration changes. Dual-mode cellphones enable users to continue their

conversations as they move between an outside cellular service and an internal Wi-Fi network, so that it is no longer necessary to carry both a desktop phone and a cellphone. Maintenance becomes simpler as there are fewer devices to oversee.

Skype, which originally marketed itself as a service among friends, has begun to cater to businesses, providing free-of-charge connections between any users on the Skype network and connecting to and from ordinary PSTN telephones for a charge.

In the United States the Social Security Administration (SSA) is converting its field offices of 63,000 workers from traditional phone installations to a VoIP infrastructure carried over its existing data network.

Benefits

Operational cost

VoIP can be a benefit for reducing communication and infrastructure costs. Examples include:

- Routing phone calls over existing data networks to avoid the need for separate voice and data networks.
- Conference calling, IVR, call forwarding, automatic redial, and caller ID features that traditional telecommunication companies (telcos) normally charge extra for, are available free of charge from open source VoIP implementations.

Flexibility

VoIP can facilitate tasks and provide services that may be more difficult to implement using the PSTN. Examples include:

- The ability to transmit more than one telephone call over a single broadband connection.
- Secure calls using standardized protocols (such as Secure Real-time Transport Protocol). Most of the difficulties of creating a secure telephone connection over traditional phone lines, such as digitizing and digital transmission, are already in place with VoIP. It is only necessary to encrypt and authenticate the existing data stream.
- Location independence. Only a sufficiently fast and stable Internet connection is needed to get a connection from anywhere to a VoIP provider.
- Integration with other services available over the Internet, including video conversation, message or data file exchange during the conversation, audio conferencing, managing address books, and passing information about whether other people are available to interested parties.
- Unified Communications, the integration of VoIP with other business systems including E-mail, Customer Relationship Management (CRM), and Web systems.

Challenges

Quality of service

Communication on the IP network is inherently less reliable in contrast to the circuit-switched public telephone network, as it does not provide a network-based mechanism to ensure that data packets are not lost, or delivered in sequential order. It is a best-effort network without fundamental Quality of Service (QoS) guarantees. Therefore, VoIP implementations may face problems mitigating latency and jitter.

By default, IP routers handle traffic on a first-come, first-served basis. Routers on high volume traffic links may introduce latency that exceeds permissible thresholds for VoIP. Fixed delays cannot be controlled, as they are caused by the physical distance the packets travel; however, latency can be minimized by marking voice packets as being delay-sensitive with methods such as DiffServ.

A VoIP packet usually has to wait for the current packet to finish transmission, although it is possible to preempt (abort) a less important packet in mid-transmission, although this is not commonly done, especially on high-speed links where transmission times are short even for maximum-sized packets. An alternative to preemption on slower links, such as dialup and DSL, is to reduce the maximum transmission time by reducing the maximum transmission unit. But every packet must contain protocol headers, so this increases relative header overhead on every link along the user's Internet paths, not just the bottleneck (usually Internet access) link.

ADSL modems provide Ethernet (or Ethernet over USB) connections to local equipment, but inside they are actually ATM modems. They use AAL5 to segment each Ethernet packet into a series of 53-byte ATM cells for transmission and reassemble them back into Ethernet packets at the receiver. A virtual circuit identifier (VCI) is part of the 5-byte header on every ATM cell, so the transmitter can multiplex the active virtual circuits (VCs) in any arbitrary order. Cells from the *same* VC are always sent sequentially.

However, the great majority of DSL providers use only one VC for each customer, even those with bundled VoIP service. Every Ethernet packet must be completely transmitted before another can begin. If a second PVC were established, given high priority and reserved for VoIP, then a low priority data packet could be suspended in mid-transmission and a VoIP packet sent right away on the high priority VC. Then the link would pick up the low priority VC where it left off. Because ATM links are multiplexed on a cell-by-cell basis, a high priority packet would have to wait at most 53 byte times to begin transmission. There would be no need to reduce the interface MTU and accept the resulting increase in higher layer protocol overhead, and no need to abort a low priority packet and resend it later.

ATM has substantial header overhead: 5/53 = 9.4%, roughly twice the total header overhead of a 1500 byte TCP/IP Ethernet packet (with TCP timestamps). This "ATM

tax" is incurred by every DSL user whether or not he takes advantage of multiple virtual circuits - and few can.

ATM's potential for latency reduction is greatest on slow links, because worst-case latency decreases with increasing link speed. A full-size (1500 byte) Ethernet frame takes 94 ms to transmit at 128 kb/s but only 8 ms at 1.5 Mb/s. If this is the bottleneck link, this latency is probably small enough to ensure good VoIP performance without MTU reductions or multiple ATM PVCs. The latest generations of DSL, VDSL and VDSL2, carry Ethernet without intermediate ATM/AAL5 layers, and they generally support IEEE 802.1p priority tagging so that VoIP can be queued ahead of less time-critical traffic.

Voice, and all other data, travels in packets over IP networks with fixed maximum capacity. This system may be more prone to congestion and DoS attacks than traditional circuit switched systems; a circuit switched system of insufficient capacity will refuse new connections while carrying the remainder without impairment, while the quality of real-time data such as telephone conversations on packet-switched networks degrades dramatically.

Fixed delays cannot be controlled as they are caused by the physical distance the packets travel. They are especially problematic when satellite circuits are involved because of the long distance to a geostationary satellite and back; delays of 400–600 ms are typical.

When the load on a link grows so quickly that its switches experience queue overflows, congestion results and data packets are lost. This signals a transport protocol like TCP to reduce its transmission rate to alleviate the congestion. But VoIP usually uses UDP not TCP because recovering from congestion through retransmission usually entails too much latency. So QoS mechanisms can avoid the undesirable loss of VoIP packets by immediately transmitting them ahead of any queued bulk traffic on the same link, even when that bulk traffic queue is overflowing.

The receiver must resequence IP packets that arrive out of order and recover gracefully when packets arrive too late or not at all. Jitter results from the rapid and random (i.e., unpredictable) changes in queue lengths along a given Internet path due to competition from other users for the same transmission links. VoIP receivers counter jitter by storing incoming packets briefly in a "de-jitter" or "playout" buffer, deliberately increasing latency to improve the chance that each packet will be on hand when it is time for the voice engine to play it. The added delay is thus a compromise between excessive latency and excessive dropout, i.e., momentary audio interruptions.

Although jitter is a random variable, it is the sum of several other random variables that are at least somewhat independent: the individual queuing delays of the routers along the Internet path in question. Thus according to the central limit theorem, we can model jitter as a gaussian random variable. This suggests continually estimating the mean delay and its standard deviation and setting the playout delay so that only packets delayed more than several standard deviations above the mean will arrive too late to be useful. In practice, however, the variance in latency of many Internet paths is dominated by a small

number (often one) of relatively slow and congested "bottleneck" links. Most Internet backbone links are now so fast (e.g. 10 Gb/s) that their delays are dominated by the transmission medium (e.g. optical fiber) and the routers driving them do not have enough buffering for queuing delays to be significant.

It has been suggested to rely on the packetized nature of media in VoIP communications and transmit the stream of packets from the source phone to the destination phone simultaneously across different routes (multi-path routing). In such a way, temporary failures have less impact on the communication quality. In capillary routing it has been suggested to use at the packet level Fountain codes or particularly raptor codes for transmitting extra redundant packets making the communication more reliable.

A number of protocols have been defined to support the reporting of QoS/QoE for VoIP calls. These include RTCP Extended Report (RFC 3611), SIP RTCP Summary Reports, H.460.9 Annex B (for H.323), H.248.30 and MGCP extensions. The RFC 3611 VoIP Metrics block is generated by an IP phone or gateway during a live call and contains information on packet loss rate, packet discard rate (because of jitter), packet loss/discard burst metrics (burst length/density, gap length/density), network delay, end system delay, signal / noise / echo level, Mean Opinion Scores (MOS) and R factors and configuration information related to the jitter buffer.

RFC 3611 VoIP metrics reports are exchanged between IP endpoints on an occasional basis during a call, and an end of call message sent via SIP RTCP Summary Report or one of the other signaling protocol extensions. RFC 3611 VoIP metrics reports are intended to support real time feedback related to QoS problems, the exchange of information between the endpoints for improved call quality calculation and a variety of other applications.

Layer-2 quality of service

A number of protocols that deal with the data link layer and physical layer include quality-of-service mechanisms that can be used to ensure that applications like VoIP work well even in congested scenarios. Some examples include:

- IEEE 802.11e is an approved amendment to the IEEE 802.11 standard that defines a set of quality-of-service enhancements for wireless LAN applications through modifications to the Media Access Control (MAC) layer. The standard is considered of critical importance for delay-sensitive applications, such as Voice over Wireless IP.
- IEEE 802.1p defines 8 different classes of service (including one dedicated to voice) for traffic on layer-2 wired Ethernet.
- The ITU-T G.hn standard, which provides a way to create a high-speed (up to 1 gigabit per second) Local area network using existing home wiring (power lines, phone lines and coaxial cables). G.hn provides QoS by means of "Contention-Free Transmission Opportunities" (CFTXOPs) which are allocated to flows (such

as a VoIP call) which require QoS and which have negotiated a "contract" with the network controller.

Susceptibility to power failure

Telephones for traditional residential analog service are usually connected directly to telephone company phone lines which provide direct current to power most basic analog handsets independently of locally available power.

IP Phones and VoIP telephone adapters connect to routers or cable modems which typically depend on the availability of mains electricity or locally generated power. Some VoIP service providers use customer premise equipment (e.g., cablemodems) with battery-backed power supplies to assure uninterrupted service for up to several hours in case of local power failures. Such battery-backed devices typically are designed for use with analog handsets.

Some VoIP service providers implement services to route calls to other telephone services of the subscriber, such a cellular phone, in the event that the customer's network device is inaccessible to terminate the call.

The susceptibility of phone service to power failures is a common problem even with traditional analog service in areas where many customers purchase modern telephone units that operate with wireless handsets to a base station, or that have other modern phone features, such as built-in voicemail or phone book features.

Emergency calls

The nature of IP makes it difficult to locate network users geographically. Emergency calls, therefore, cannot easily be routed to a nearby call center. Sometimes, VoIP systems may route emergency calls to a non-emergency phone line at the intended department. In the United States, at least one major police department has strongly objected to this practice as potentially endangering the public.

A fixed line phone has a direct relationship between a telephone number and a physical location. If an emergency call comes from that number, then the physical location is known.

In the IP world, it is not so simple. A broadband provider may know the location where the wires terminate, but this does not necessarily allow the mapping of an IP address to that location. IP addresses are often dynamically assigned, so the ISP may allocate an address for online access, or at the time a broadband router is engaged. The ISP recognizes individual IP addresses, but does not necessarily know to which physical location it corresponds. The broadband service provider knows the physical location, but is not necessarily tracking the IP addresses in use.

There are more complications since IP allows a great deal of mobility. For example, a broadband connection can be used to dial a virtual private network that is employer-owned. When this is done, the IP address being used will belong to the range of the employer, rather than the address of the ISP, so this could be many kilometres away or even in another country. To provide another example: if mobile data is used, e.g., a 3G mobile handset or USB wireless broadband adapter, then the IP address has no relationship with any physical location, since a mobile user could be anywhere that there is network coverage, even roaming via another cellular company.

In short, there is no relationship between IP address and physical location, so the address itself reveals no useful information for the emergency services.

At the VoIP level, a phone or gateway may identify itself with a SIP registrar by using a username and password. So in this case, the Internet Telephony Service Provider (ITSP) knows that a particular user is online, and can relate a specific telephone number to the user. However, it does not recognize how that IP traffic was engaged. Since the IP address itself does not necessarily provide location information presently, today a "best efforts" approach is to use an available database to find that user and the physical address the user chose to associate with that telephone number—clearly an imperfect solution.

VoIP Enhanced 911 (E911) is a method by which VoIP providers in the United States support emergency services. The VoIP E911 emergency-calling system associates a physical address with the calling party's telephone number as required by the Wireless Communications and Public Safety Act of 1999. All VoIP providers that provide access to the public switched telephone network are required to implement E911, a service for which the subscriber may be charged. Participation in E911 is not required and customers may opt-out of E911 service.

One shortcoming of VoIP E911 is that the emergency system is based on a static table lookup. Unlike in cellular phones, where the location of an E911 call can be traced using Assisted GPS or other methods, the VoIP E911 information is only accurate so long as subscribers are diligent in keeping their emergency address information up-to-date. In the United States, the Wireless Communications and Public Safety Act of 1999 leaves the burden of responsibility upon the subscribers and not the service providers to keep their emergency information up to date.

Lack of redundancy

With the current separation of the Internet and the PSTN, a certain amount of redundancy is provided. An Internet outage does not necessarily mean that a voice communication outage will occur simultaneously, allowing individuals to call for emergency services and many businesses to continue to operate normally. In situations where telephone services become completely reliant on the Internet infrastructure, a single-point failure can isolate communities from all communication, including Enhanced 911 and equivalent services in other locales. However, the internet as designed by DARPA in the early 1980s was specifically designed to be fault tolerant under adverse conditions. Even during the 9/11

attacks on the World Trade Centers the internet routed data around the failed nodes that were housed in or near the towers. So single point failures while possible in some geographic areas are not the norm for the internet as a whole.

Number portability

Local number portability (LNP) and Mobile number portability (MNP) also impact VoIP business. In November 2007, the Federal Communications Commission in the United States released an order extending number portability obligations to interconnected VoIP providers and carriers that support VoIP providers. Number portability is a service that allows a subscriber to select a new telephone carrier without requiring a new number to be issued. Typically, it is the responsibility of the former carrier to "map" the old number to the undisclosed number assigned by the new carrier. This is achieved by maintaining a database of numbers. A dialed number is initially received by the original carrier and quickly rerouted to the new carrier. Multiple porting references must be maintained even if the subscriber returns to the original carrier. The FCC mandates carrier compliance with these consumer-protection stipulations.

A voice call originating in the VoIP environment also faces challenges to reach its destination if the number is routed to a mobile phone number on a traditional mobile carrier. VoIP has been identified in the past as a Least Cost Routing (LCR) system, which is based on checking the destination of each telephone call as it is made, and then sending the call via the network that will cost the customer the least. This rating is subject to some debate given the complexity of call routing created by number portability. With GSM number portability now in place, LCR providers can no longer rely on using the network root prefix to determine how to route a call. Instead, they must now determine the actual network of every number before routing the call.

Therefore, VoIP solutions also need to handle MNP when routing a voice call. In countries without a central database, like the UK, it might be necessary to query the GSM network about which home network a mobile phone number belongs to. As the popularity of VoIP increases in the enterprise markets because of least cost routing options, it needs to provide a certain level of reliability when handling calls.

MNP checks are important to assure that this quality of service is met. By handling MNP lookups before routing a call and by assuring that the voice call will actually work, VoIP service providers are able to offer business subscribers the level of reliability they require.

PSTN integration

E.164 is a global FGFnumbering standard for both the PSTN and PLMN. Most VoIP implementations support E.164 to allow calls to be routed to and from VoIP subscribers and the PSTN/PLMN. VoIP implementations can also allow other identification techniques to be used. For example, Skype allows subscribers to choose "Skype names" (usernames) whereas SIP implementations can use URIs similar to email addresses. Often

VoIP implementations employ methods of translating non-E.164 identifiers to E.164 numbers and vice-versa, such as the Skype-In service provided by Skype and the ENUM service in IMS and SIP.

Echo can also be an issue for PSTN integration. Common causes of echo include impedance mismatches in analog circuitry and acoustic coupling of the transmit and receive signal at the receiving end.

Security

VoIP telephone systems are susceptible to attacks as are any internet-connected devices. This means that hackers who know about these vulnerabilities (such as insecure passwords) can institute denial-of-service attacks, harvest customer data, record conversations and break into voice mailboxes.

Another challenge is routing VoIP traffic through firewalls and network address translators. Private Session Border Controllers are used along with firewalls to enable VoIP calls to and from protected networks. For example, Skype uses a proprietary protocol to route calls through other Skype peers on the network, allowing it to traverse symmetric NATs and firewalls. Other methods to traverse NATs involve using protocols such as STUN or ICE.

Many consumer VoIP solutions do not support encryption, although having a secure phone is much easier to implement with VoIP than traditional phone lines. As a result, it is relatively easy to eavesdrop on VoIP calls and even change their content. An attacker with a packet sniffer could intercept your VoIP calls if you are not on a secure VLAN. However, physical security of the switches within an enterprise and the facility security provided by ISPs make packet capture less of a problem than originally foreseen. Further research has shown that tapping into a fiber optic network without detection is difficult if not impossible. This means that once a voice packet is within the internet backbone it is relatively safe from interception.

There are open source solutions, such as Wireshark, that facilitate sniffing of VoIP conversations. A modicum of security is afforded by patented audio codecs in proprietary implementations that are not easily available for open source applications; however, such security through obscurity has not proven effective in other fields. Some vendors also use compression, which may make eavesdropping more difficult. However, real security requires encryption and cryptographic authentication which are not widely supported at a consumer level. The existing security standard Secure Real-time Transport Protocol (SRTP) and the new ZRTP protocol are available on Analog Telephone Adapters (ATAs) as well as various softphones. It is possible to use IPsec to secure P2P VoIP by using opportunistic encryption. Skype does not use SRTP, but uses encryption which is transparent to the Skype provider. In 2005, Skype invited a researcher, Dr Tom Berson, to assess the security of the Skype software, and his conclusions are available in a published report.

The Voice VPN solution provides secure voice for enterprise VoIP networks by applying IPSec encryption to the digitized voice stream.

Securing VoIP

To prevent the above security concerns government and military organizations are using Voice over Secure IP (VoSIP), Secure Voice over IP (SVoIP), and Secure Voice over Secure IP (SVoSIP) to protect confidential and classified VoIP communications. Secure Voice over IP is accomplished by encrypting VoIP with Type 1 encryption. Secure Voice over Secure IP is accomplished by using Type 1 encryption on a classified network, like SIPRNet. Public Secure VoIP is also available with free GNU programs.

Caller ID

Caller ID support among VoIP providers varies, although the majority of VoIP providers now offer full Caller ID with name on outgoing calls.

In a few cases, VoIP providers may allow a caller to spoof the Caller ID information, potentially making calls appear as though they are from a number that does not belong to the caller Business grade VoIP equipment and software often makes it easy to modify caller ID information. Although this can provide many businesses great flexibility, it is also open to abuse.

The "Truth in Caller ID Act" has been in preparation in the US Congress since 2006, but as of January 2009 still has not been enacted. This bill proposes to make it a crime in the United States to "knowingly transmit misleading or inaccurate caller identification information with the intent to defraud, cause harm, or wrongfully obtain anything of value ..."

Compatibility with traditional analog telephone sets

Some analog telephone adapters do not decode pulse dialing from older phones. They may only work with push-button telephones using the touch-tone system. The VoIP user may use a pulse-to-tone converter, if needed.

Fax handling

Support for sending faxes over VoIP implementations is still limited. The existing voice codecs are not designed for fax transmission; they are designed to digitize an analog representation of a human voice efficiently. However, the inefficiency of digitizing an analog representation (modem signal) of a digital representation (a document image) of analog data (an original document) more than negates any bandwidth advantage of VoIP. In other words, the fax "sounds" simply do not fit in the VoIP channel. An alternative IP-based solution for delivering fax-over-IP called T.38 is available.

The T.38 protocol is designed to compensate for the differences between traditional packet-less communications over analog lines and packet based transmissions which are the basis for IP communications. The fax machine could be a traditional fax machine connected to the PSTN, or an ATA box (or similar). It could be a fax machine with an RJ-45 connector plugged straight into an IP network, or it could be a computer pretending to be a fax machine. Originally, T.38 was designed to use UDP and TCP transmission methods across an IP network. TCP is better suited for use between two IP devices. However, older fax machines, connected to an analog system, benefit from UDP near real-time characteristics due to the "no recovery rule" when a UDP packet is lost or an error occurs during transmission. UDP transmissions are preferred as they do not require testing for dropped packets and as such since each T.38 packet transmission includes a majority of the data sent in the prior packet, a T.38 termination point has a higher degree of success in re-assembling the fax transmission back into its original form for interpretation by the end device. This in an attempt to overcome the obstacles of simulating real time transmissions using packet based protocol.

There have been updated versions of T.30 to resolve the fax over IP issues, which is the core fax protocol. Some newer high end fax machines have T.38 built-in capabilities which allow the user to plug right into the network and transmit/receive faxes in native T.38 like the Ricoh 4410NF Fax Machine. A unique feature of T.38 is that each packet contains a portion of the main data sent in the previous packet. With T.38, two successive lost packets are needed to actually lose any data. The data you lose will only be a small piece, but with the right settings and error correction mode, there is an increased likelihood that you will receive enough of the transmission to satisfy the requirements of the fax machine for output of the sent document.

Support for other telephony devices

Another challenge for VoIP implementations is the proper handling of outgoing calls from other telephony devices such as Digital Video RecordersDVR boxes, satellite television receivers, alarm systems, conventional modems and other similar devices that depend on access to a PSTN telephone line for some or all of their functionality.

These types of calls sometimes complete without any problems, but in other cases they fail. If VoIP and cellular substitution becomes very popular, some ancillary equipment makers may be forced to redesign equipment, because it would no longer be possible to assume a conventional PSTN telephone line would be available in consumer's homes.

Legal issues

As the popularity of VoIP grows, and PSTN users switch to VoIP in increasing numbers, governments are becoming more interested in regulating VoIP in a manner similar to PSTN services.

Another legal issue that the US Congress is debating concerns changes to the Foreign Intelligence Surveillance Act. The issue in question is calls between Americans and

foreigners. The National Security Agency (NSA) is not authorized to tap Americans' conversations without a warrant—but the Internet, and specifically VoIP does not draw as clear a line to the location of a caller or a call's recipient as the traditional phone system does. As VoIP's low cost and flexibility convinces more and more organizations to adopt the technology, the surveillance for law enforcement agencies becomes more difficult. VoIP technology has also increased security concerns because VoIP and similar technologies have made it more difficult for the government to determine where a target is physically located when communications are being intercepted, and that creates a whole set of new legal challenges.

In the US, the Federal Communications Commission now requires all interconnected VoIP service providers to comply with requirements comparable to those for traditional telecommunications service providers. VoIP operators in the US are required to support local number portability; make service accessible to people with disabilities; pay regulatory fees, universal service contributions, and other mandated payments; and enable law enforcement authorities to conduct surveillance pursuant to the Communications Assistance for Law Enforcement Act (CALEA). "Interconnected" VoIP operators also must provide Enhanced 911 service, disclose any limitations on their E-911 functionality to their consumers, and obtain affirmative acknowledgements of these disclosures from all consumers. VoIP operators also receive the benefit of certain US telecommunications regulations, including an entitlement to interconnection and exchange of traffic with incumbent local exchange carriers via wholesale carriers. Providers of "nomadic" VoIP service—those who are unable to determine the location of their users—are exempt from state telecommunications regulation.

Throughout the developing world, countries where regulation is weak or captured by the dominant operator, restrictions on the use of VoIP are imposed, including in Panama where VoIP is taxed, Guyana where VoIP is prohibited and India where its retail commercial sales is allowed but only for long distance service. In Ethiopia, where the government is monopolizing telecommunication service, it is a criminal offense to offer services using VoIP. The country has installed firewalls to prevent international calls being made using VoIP. These measures were taken after the popularity of VoIP reduced the income generated by the state owned telecommunication company.

In the European Union, the treatment of VoIP service providers is a decision for each Member State's national telecoms regulator, which must use competition law to define relevant national markets and then determine whether any service provider on those national markets has "significant market power" (and so should be subject to certain obligations). A general distinction is usually made between VoIP services that function over managed networks (via broadband connections) and VoIP services that function over unmanaged networks (essentially, the Internet).

VoIP services that function over managed networks are often considered to be a viable substitute for PSTN telephone services (despite the problems of power outages and lack of geographical information); as a result, major operators that provide these services (in

practice, incumbent operators) may find themselves bound by obligations of price control or accounting separation.

VoIP services that function over unmanaged networks are often considered to be too poor in quality to be a viable substitute for PSTN services; as a result, they may be provided without any specific obligations, even if a service provider has "significant market power".

The relevant EU Directive is not clearly drafted concerning obligations which can exist independently of market power (e.g., the obligation to offer access to emergency calls), and it is impossible to say definitively whether VoIP service providers of either type are bound by them. A review of the EU Directive is under way and should be complete by 2007.

In India, it is legal to use VoIP, but it is illegal to have VoIP gateways inside India. This effectively means that people who have PCs can use them to make a VoIP call to any number, but if the remote side is a normal phone, the gateway that converts the VoIP call to a POTS call should not be inside India.

In the UAE and Oman it is illegal to use any form of VoIP, to the extent that Web sites of Skype and Gizmo5 are blocked. Providing or using VoIP services is illegal in Oman. Those who violate the law stand to be fined 50,000 Omani Rial (about 130,317 US dollars) or spend two years in jail or both. In 2009, police in Oman have raided 121 internet cafes throughout the country and arrested 212 people for using/providing VoIP services.

In the Republic of Korea, only providers registered with the government are authorized to offer VoIP services. Unlike many VoIP providers, most of whom offer flat rates, Korean VoIP services are generally metered and charged at rates similar to terrestrial calling. Foreign VoIP providers encounter high barriers to government registration. This issue came to a head in 2006 when Internet service providers providing personal Internet services by contract to United States Forces Korea members residing on USFK bases threatened to block off access to VoIP services used by USFK members as an economical way to keep in contact with their families in the United States, on the grounds that the service members' VoIP providers were not registered. A compromise was reached between USFK and Korean telecommunications officials in January 2007, wherein USFK service members arriving in Korea before June 1, 2007, and subscribing to the ISP services provided on base may continue to use their US-based VoIP subscription, but later arrivals must use a Korean-based VoIP provider, which by contract will offer pricing similar to the flat rates offered by US VoIP providers.

International VoIP implementation

IP telephony in Japan

In Japan, IP telephony is regarded as a service applied by VoIP technology to the whole or a part of the telephone line. As of 2003, IP telephony services have been assigned telephone numbers. IP telephony services also often include videophone/video conferencing services. According to the Telecommunication Business Law, the service category for IP telephony also implies the service provided via Internet, which is not assigned any telephone number.

IP telephony is basically regulated by Ministry of Internal Affairs and Communications (MIC) as a telecommunication service. The operators have to disclose necessary information on its quality, etc., prior to making contracts with customers, and have an obligation to respond to their complaints cordially.

Many Japanese Internet service providers (ISP) are including IP telephony services. An ISP who also provides IP telephony service is known as a "ITSP (Internet Telephony Service Provider)". Recently, the competition among ITSPs has been activated, by option or set sales, in connection with ADSL or FTTH services.

The tariff system normally applied to Japanese IP telephony is described below;

- A call between IP telephony subscribers, limited to the same group, is usually free of charge.
- A call from IP telephony subscribers to a fixed line or PHS is usually a uniformly fixed rate all over the country.

Between ITSPs, the interconnection is mostly maintained at VoIP level.

- Where the IP telephony is assigned normal telephone number (0AB-J), the condition for its interconnection is considered same as normal telephony.
- Where the IP telephony is assigned specific telephone number (050), the condition for its interconnection is described below;
    - Interconnection is sometimes charged. (Sometimes, it is free of charge.) In case of free-of-charge, mostly, communication traffic is exchanged via a P2P connection with the same VoIP standard. Otherwise, certain conversions are needed at the point of the VoIP gateway which incurs operating costs.

Since September 2002, the MIC has assigned IP telephony telephone numbers on the condition that the service falls into certain required categories of quality.

High-quality IP telephony is assigned a telephone number, normally starting with the digits 050. When VoIP quality is so high that a customer has difficulty telling the difference between it and a normal telephone, and when the provider relates its number

with a location and provides the connection with emergency call capabilities, the provider is allowed to assign a normal telephone number, which is a so-called "0AB-J" number.

Voice over IP can be used together with static IP addresses so that one can talk to any computer just the way one uses internet, but instead he can access IP-address as definitive unique 'Internet VoIP'-phone number...

Historical milestones

- 1974 – The Institute of Electrical and Electronic Engineers (IEEE) published a paper titled "A Protocol for Packet Network Interconnection."
- 1981 – IPv4 is described in RFC 791.
- 1985 – The National Science Foundation commissions the creation of NSFNET.
- 1995 – VocalTec releases the first commercial Internet phone software.
- 1996 –
    - ITU-T begins development of standards for the transmission and signaling of voice communications over Internet Protocol networks with the H.323 standard.
    - US telecommunication companies petition the US Congress to ban Internet phone technology.
- 1997 – Level 3 began development of its first softswitch, a term they coined in 1998.
- 1999 –
    - The Session Initiation Protocol (SIP) specification RFC 2543 is released.
    - Mark Spencer of Digium develops the first open source Private branch exchange (PBX) software (Asterisk).
- 2004 – Commercial VoIP service providers proliferate.