Digitized Speech

Much of the information transmitted over wide-area networks is digitized audio and video. This lecture describes how speech signals are digitized for transmission over telephony networks. Similar techniques are used for digital transmission and recording of music, documents, photographs and video.
After this lecture you should be able to: decide when it is worthwhile to digitize speech; solve problems involving: the frequencies of the desired, sampling and alias signals; sampling rate and bandwidth; bits per sample and quantization SNR; sampling rate, bits/sample and data rate. You should be able to explain how companding increases the average quantization SNR.

Introduction

Telephony is the transmission of speech over distance. The modern PSTN converts the analog POTS speech signal to and from a digital signal at the CO. This digital signal is then transmitted over digital trunks between COs. This lecture describes the conversion of speech between analog to digital forms.

The use of digitized speech transmission actually predates the development of data networks because there are many advantages to transmitting speech as bits[1]:

- Digital circuits can operate at lower signal-to-noise ratios (SNRs) because only two (or a few) levels need to be distinguished. This increases the distance between amplifiers or repeaters.
- Additive noise is eliminated when a digital signal is "regenerated" (put through a receiver-transmitter pair).
- Digital transmissions systems are less expensive because they do not need to be designed for low distortion and can be implemented using digital IC technology (with smaller die area than analog circuits).
- Signalling information can be carried out-of-band but on the same transmission system as the speech. This simplifies the design of the system by eliminating the need for tone detectors and generators (e.g. for call progress tones and DTMF) at intermediate points.
- The quality of transmission can be monitored by measuring the bit or frame error rate (BER or FER).
- Data and telephony services can be carried by the same transmission system.
- Transmissions can be easily and effectively encrypted.

[1] Bellamy, *Digital Telephony*, 2000

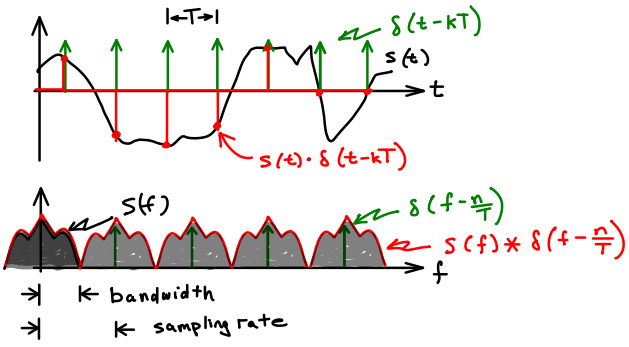
Today analog transmission of speech is limited to legacy systems that have a large user base which would mean a high cost of conversion (e.g. AM radio broadcasting) and to short-range point-to-point systems (e.g. intercoms) where none of the above advantages apply.

Exercise 1: Give some examples of legacy analog speech communications and very simple analog speech communication systems.

Sampling

Nyquist showed that a low-pass signal can be reconstructed exactly by low-pass filtering samples of the signal taken at a frequency that is at least twice the bandwidth of the signal.

In the frequency domain this can be shown by considering sampling as the equivalent of multiplying the signal with a periodic sequences of impulses as shown below. Sampling causes the two-sided spectrum of the sampled signal to be replicated in frequency with a spacing equal to the impulse (sampling) frequency. The replicated spectra of the sampled signal will not overlap as long as the sampling rate is *twice* the bandwidth.



If the sampling rate is not high enough, the high-frequency portion of the spectrum (from the negative frequency components) falls into the signal spectrum. This results in a type of distortion called "aliasing."

Exercise 2: A 5 kHz signal is sampled at 8 kHz. What are the positive and negative frequency components of the 5 kHz signal before sampling? What is the frequency of the aliased component falling into the 0-4 kHz range?

This frequency range (and an SNR of more than about 30dB) results in what is called "toll quality" speech – speech that has long been considered "good enough" for most customers. However, higher bandwidths and SNRs result in subjectively better quality. So-called wideband codecs are coming into wider use in internet-based telephony where there is no need to remain compatible with the sampling rates used by legacy telephony systems.

Anti-Aliasing Filter

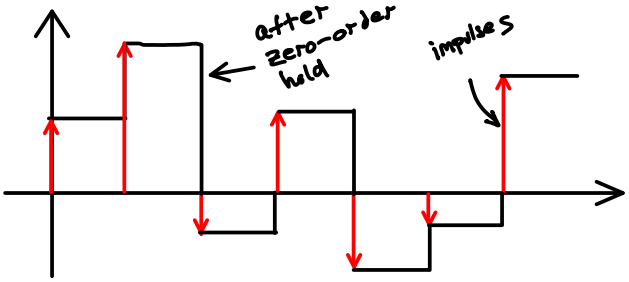
To avoid aliasing, the speech signal must be low-pass filtered to remove frequency components at frequencies higher than half the sampling rate.

The standard for telephony is to filter out frequencies below 300 Hz and above 3.4 kHz and to use a sampling rate of 8 kHz.

Reconstruction Filter

The sampled signal must be low-pass filtered to recover the analog signal.

The D/A converter generates a continuous voltage that steps between samples values rather than a sequence of impulses. This can be considered as a "zero-order hold" filtering operation where the impulse response of the filter is a pulse shape with a width equal to the sampling period. The effect in the frequency domain is a low-pass sinc(f) filtering operation. To avoid distortion, the spectrum of the signal needs to be corrected by applying "$x/\sin(x)$" correction that boosts higher-frequency components. Most telephony codecs include this function.



Quantization Noise

In addition to sampling the signal at discrete time intervals, each sample voltage also needs to be converted to a number. Each number corresponds to a different voltage and therefore the signal must be "rounded off" to the nearest number. This quantization operation is equivalent to adding a "quantization noise" signal to the un-quantized signal.

The samples are quantized to binary numbers. This means the number of voltage steps is a power of two. Increasing the number of bits in the numbers representing the sample value reduces the voltage difference between quantized values.

Adding one bit of resolution halves the size of the quantization steps and thus reduces the quantization noise voltage by half and the quantization noise power by a quarter (6 dB).

The actual ratio of the signal power to quantization noise power (quantization SNR) is difficult to calculate because it depends on the probability distribution of the signal and because there is a trade-off between clipping distortion and quantization noise – increasing the signal level increases signal power but also increases clipping noise. In addition, companding (described below) will affect the quantization SNR.

However, for a signal that is uniformly distributed over the possible values, the quantization SNR in dB is simply $6b$ where b is the number of bits. This is a reasonable approximation for other situations involving uniform quantization.

Exercise 3: What is the quantization SNR for a sawtooth wave varying from 0 to 1V if a 7 bit A/D converter is used with an input range of 0 to 2 V?

Companding

The quantization noise power is a function of the step size while the signal power is a function of the signal

voltage. Thus the quantization SNR is higher at higher input levels and lower at lower input levels. Companding, a combination of the words compressing and expanding, is a way to increase the average quantization SNR by effectively using small quantization steps for low levels and large quantization steps at high levels. Companding is defined by a non-linear conversion before the A/D converter than provides higher gain for low signal levels and less gain for high signal levels.

PCM Standards

ITU-T standard G.711 defines defines the sampling rate (8 kHz $\pm$50ppm), signal bandwidth (300 to 3400 Hz), and two types of companding to be used for digital telephony (μ-law and A-law).

The μ-law companding curve is primarily used in North America and A-law is primarily used in Europe. Most hardware and software can be configured to use either.

For historical reasons this method of digitizing signals is called PCM (pulse code modulation). The hardware that converts the analog signal to/from digital format is often called a codec (coder-decoder) and is often integrated into the SLICs in the line card.

Exercise 4: If the sampling rate is 8 kHz and there are 8 bits per sample, that is the data rate in each direction? How many bytes per minute are transmitted for a two-way connection?

Speech Coding

Some systems use lossy compression of speech to reduce the data rate that needs to be transmitted. Compression is used by all cellular systems due to the limited data rates available. Compression is sometimes also used for Internet telephony and speech storage.

There are a wide range of speech codecs with data rates down to about 300 bps but more common rates are around 8 kb/s for toll-quality (e.g. ITU-T G.729). Lower data rates tend to have lower quality, possibly losing speaker recognition and more complex signal processing requirements.

Packetized Speech

Today much of the speech traffic is carried over data networks. When the data network protocol is IP this is called Voice over IP (VoIP). Additional steps are often needed to ensure good quality speech when data is to be transmitted over such data networks:

Packetization: Since speech is continuous it must be broken up into packets. The packet length is a compromise between efficiency (longer packets create less overhead) and delay (shorter packets result in less delay). A typical packet length is 20 ms.

Buffering: Packets travelling across a network will be delayed, may arrive in a different order than they were send and may be lost altogether. At receiver packets are queued up in a reassembly or "jitter buffer" to compensate for variable propagation delays across the network.

Echo Cancellation: Most telephone sets and headsets cause some of the speaker output signal to be picked up by the microphone. If the end-to-end propagation delays are short this fed-back signal is indistinguishable from sidetone.

However, when the delay is long enough to be noticeable, the effect is a distracting echo. Because of the significant delays caused by packetization and network delays, echo cancellation is particularly important for VoIP. An echo canceler compares the outgoing and incoming signals and subtracts a delayed and scaled portion of the outgoing signal from the received signal to cancel the echo from the remote end. The echo canceler has to adapt to varying delays (e.g. due to changing jitter buffer lengths) and echo level.

Speech Activity Detection: If we detect when a user is not talking we can transmit fewer packets and reduce the average data rate by about a half. At the remote end random noise ("comfort" noise) noise is often generated to hide the fact that no speech is being received from the other end during these silent intervals.

Loss Concealment: If a packet is lost or doesn't arrive in time to be output in order many systems use loss concealment techniques such as repeating the sounds in the previous packet.