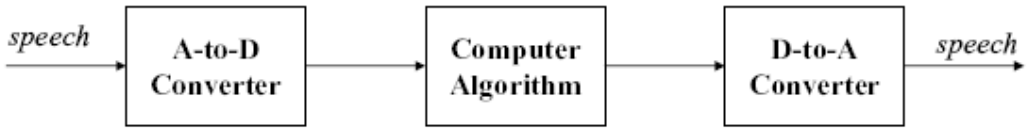
Digital Speech Processing

1.Introduction

- The frequency range of sounds perceived by humans is something between 30 Hz (hertz, cycles per second) and 20,000 Hz. Above 20K we hear nothing.
- The acoustic file need to format in which speech data can be saved. In recent years the most popular format (most frequently used format) has been Microsoft WAV format. There are many other formats(MKV, AVI.MP3). The difference is in how many samples occur per second, and exactly each sample is represented.

2.Applications of Digital Speech Processing

The first step in most applications of digital speech processing is to convert the acoustic waveform to a sequence of numbers. Most modern A-to-D converters operate by sampling at a very high rate, applying a digital low pass filter with cutoff set to preserve a prescribed bandwidth, and then reducing the sampling rate to the desired sampling rate, which can be as low as twice the cutoff frequency of the sharp-cutoff digital filter. This discrete-time representation is the starting point for most applications. From this point, other representations are obtained by digital processing



General block diagram for application of digital signal processing to speech signals.

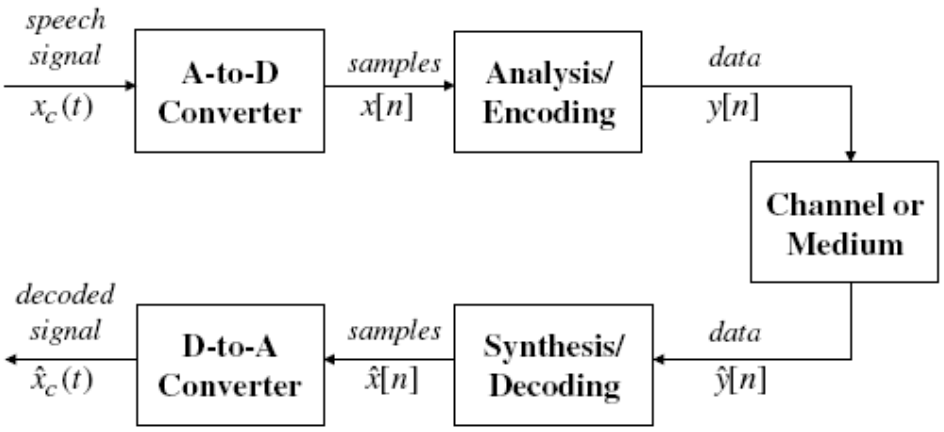
- The field of voice processing encompasses five broad technology areas, including:
- voice coding, the process of compressing the information in a voice signal so as to either transmit it or store it economically over a

channel whose bandwidth is significantly smaller than that of the uncompressed signal;

- voice synthesis, the process of creating a synthetic replica of a voice signal so as to transmit a message from a machine to a person, with the purpose of conveying the information in the message;
- speech recognition, the process of extracting the message information in a voice signal so as to control the actions of a machine in response to spoken commands;
- speaker recognition, the process of either identifying or verifying a speaker by extracting individual voice characteristics, primarily for the purpose of restricting access to information (e.g., personal/ private records), networks , or physical premises.
- • spoken language translation, the process of recognizing the speech of a person talking in one language, translating the message content to a second language, and synthesizing an appropriate message in the second language, for the purpose of providing two-way communication between people who do not speak the same language.

3.Speech Coding

- Perhaps the most widespread applications of digital speech processing technology occur in the areas of digital transmission and storage of speech signals. In these areas the centrality of the digital representation is obvious, since the goal is to compress the digital waveform representation of speech into a lower bit-rate representation. It is common to refer to this activity as “speech coding” or “speech compression.”



Speech coding block diagram — encoder and decoder.

- Figure shows a block diagram of a generic speech encoding/decoding (or compression) system. In the upper part of the figure, the A-to-D converter converts the analog speech signal $xc(t)$ to a sampled waveform representation $x[n]$. The digital signal $x[n]$ is analyzed and coded by digital computation algorithms to produce a new digital signal $y[n]$ that can be transmitted over a communication channel or stored in a storage medium as $\hat{y}[n]$. As we will see, there are a myriad of ways to do the encoding so as to reduce the data rate over that of the sampled and quantized speech waveform $x[n]$. Because the digital representation at this point is often not directly related to the sampled speech waveform, $y[n]$ and $\hat{y}[n]$ are appropriately referred to as *data signals* that represent the speech signal. The lower path in Figure shows the decoder associated with the speech coder. The received data signal $\hat{y}[n]$ is decoded using the inverse of the analysis processing, giving the sequence of samples $\hat{x}[n]$ which is then converted (using a D-to-A Converter) back to an analog signal $\hat{x}c(t)$ for human listening. The decoder is often called a *synthesizer* because it must reconstitute the speech waveform from data that may bear no direct relationship to the waveform.
- The compressed representation can be more efficiently transmitted or stored, or the bits saved can be devoted to error protection. Speech coders enable a broad range of applications including narrowband and broadband wired telephony, cellular

communications, voice over internet protocol (VoIP) (which utilizes the internet as a real-time communications medium), secure voice for privacy and encryption (for national security applications), and for storage of speech for telephone answering machines. Speech coders often utilize many aspects of both the speech production and speech perception processes, and hence may not be useful for more general audio signals such as music. Coders that are based on incorporating only aspects of sound perception generally do not achieve as much compression as those based on speech production, but they are more general and can be used for all types of audio signals. These coders are widely deployed in MP3 and for audio in digital television systems

4.Speech Recognition and Other Pattern Matching Problems

- Another large class of digital speech processing applications is concerned with the automatic extraction of information from the speech signal. Most such systems involve some sort of pattern matching. Figure shows a block diagram of a generic approach to pattern matching problems in speech processing. Such problems include the following: speech recognition, where the object is to extract the message from the speech signal.

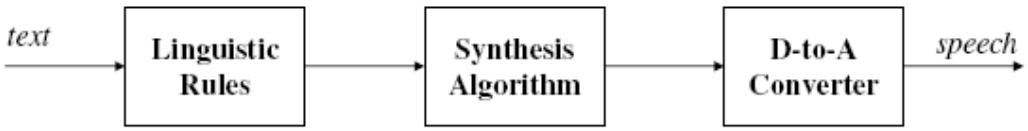
Block diagram of general pattern matching system for speech signals.

- speaker recognition, where the goal is to identify who is speaking; speaker verification, where the goal is to verify a speaker's claimed identity from analysis of their speech
- signal; word spotting, which involves monitoring a speech signal for the occurrence of specified words or phrases; and automatic indexing of speech recordings based on recognition (or spotting) of spoken keywords.

- The first block in the pattern matching system converts the analog speech waveform to digital form using an A-to-D converter. The feature analysis module converts the sampled speech signal to a set of feature vectors. Often, the same analysis techniques that are used in speech coding are also used to derive the feature vectors. The final block in the system, namely the pattern matching block, the set of feature vectors representing the speech signal with a concatenated set of stored patterns, and chooses the identity associated with the pattern which is the closest match to the set of feature vectors of the speech signal. The symbolic output consists of a set of recognized words, in the case of speech recognition, or the identity of the best matching talker, in the case of speaker recognition, or a decision as to whether to accept or reject the identity claim of a speaker in the case of speaker verification.
- Another major speech application that has long been a dream of speech researchers is *automatic language translation*. The goal of language translation systems is to convert spoken words in one language to spoken words in another language so as to facilitate natural language voice dialogues between people speaking different languages. Language translation technology requires speech synthesis systems that work in both languages, along with speech recognition (and generally natural language understanding) that also works for both languages; hence it is a very difficult task and one for which only limited progress has been made. When such systems exist, it will be possible for people speaking different languages to communicate at data rates on the order of that of printed text reading!

5.Text-to-Speech Synthesis

- For many years, scientists and engineers have studied the speech production process with the goal of building a system that can start with text and produce speech automatically. In a sense, a text-to-speech synthesizer such as depicted in Figure is a digital simulation of the entire upper part of the speech chain diagram.



Text-to-speech synthesis system block diagram.

- The input to the system is ordinary text such as an email message or an article from a newspaper or magazine. The first block in the text-to-speech synthesis system, labeled linguistic rules, has the job of converting the printed text input into a set of sounds that the machine must synthesize. The conversion from text to sounds involves a set of linguistic rules that must determine the appropriate set of sounds (perhaps including things like emphasis, pauses, rates of speaking, etc.) so that the resulting synthetic speech will express the words and intent of the text message in what passes for a natural voice that can be decoded accurately by human speech perception. This is more difficult than simply looking up the words in a pronouncing dictionary because the linguistic rules must determine how to pronounce acronyms, how to pronounce ambiguous words like *read*, *bass*, *object*, how to pronounce abbreviations like St. (street or Saint), Dr. (Doctor or drive), and how to properly pronounce proper names, specialized terms, etc.
- Once the proper pronunciation of the text has been determined, the role of the synthesis algorithm is to create the appropriate sound sequence to represent the text message in the form of speech. In essence, the synthesis algorithm must simulate the action of the vocal tract system in creating the sounds of speech. There are many procedures for assembling the speech sounds and compiling them into a proper sentence, but the most promising one today is called "unit selection and concatenation." In this method, the computer stores multiple versions of each of the basic units of speech (phones, half phones, syllables, etc.), and then decides which sequence of speech units sounds best for the particular text message that is being produced. The basic digital representation is not generally the sampled speech wave. Instead, some sort of compressed representation is normally used to
- save memory and, more importantly, to allow convenient manipulation of durations and blending of adjacent sounds. Thus, the speech synthesis algorithm would include an appropriate decoder, whose output is converted to an analog representation via the D-to-A converter. Text-to-speech synthesis systems are an essential component of modern human–machine communications systems and are used to do things like read email messages over a telephone, provide voice output from GPS systems in automobiles, provide the voices for talking agents for completion of transactions over the internet, handle call center help desks and customer care applications, serve as the voice for providing information from

handheld devices such as foreign language phrasebooks, dictionaries, crossword puzzle helpers, and as the voice of announcement machines that provide information such as stock quotes, airline schedules, updates on arrivals and departures of flights, etc. Another important application is in reading machines for the blind, where an optical character recognition system provides the text input to a speech synthesis system.

- Speech Analysis

- Short Time Energy

serves to **differentiate voiced and unvoiced sounds** in speech from silence (background signal)

$$E = \sum_{m=-\infty}^{\infty} x^2(m)$$

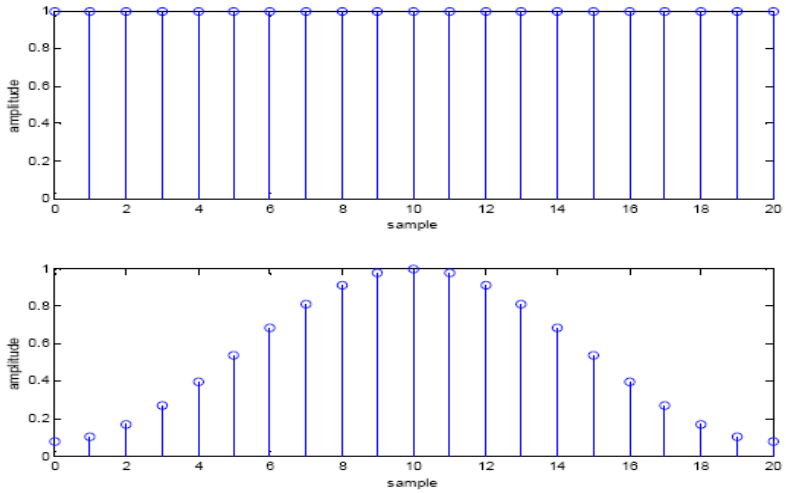
- natural definition of energy of weighted signal is:

$$E_n = \sum_{m=-\infty}^{\infty} \left[x(m)w(n-m)\right]^2 \text{ (sum or squares of portion of signal)}$$

windows

consider two windows, w(n)
– rectangular window:
• h(n)=1, 0≤n≤N-1 and 0 otherwise
– Hamming window (raised cosine window):
• h(n)=0.54-0.46 cos(2πn/(N-1)), 0≤n≤N-1 and 0 otherwise
– rectangular window gives ***equal weight*** to all N samples in the window (n,...,n-N+1)
– Hamming window gives ***most weight*** to middle samples and ***tapers off*** strongly at the beginning the end of the window.

Rectangular and Hamming Windows



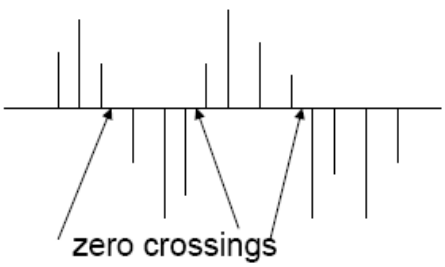
Short-Time Magnitude

- • short-time energy is very sensitive to large
- signal levels due to x2(n) terms
- – consider a new definition of 'pseudo-energy' based
- on average signal magnitude (rather than energy)

$$M_n = \sum_{m=-\infty}^{\infty} | x(m) | \, w(n-m)$$

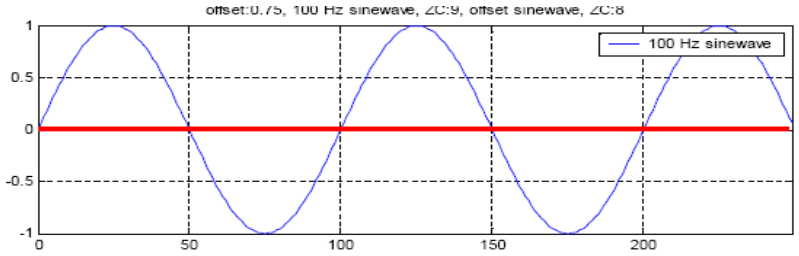
- weighted sum of magnitudes, rather than weighted sum of squares

Short-Time Average ZC Rate



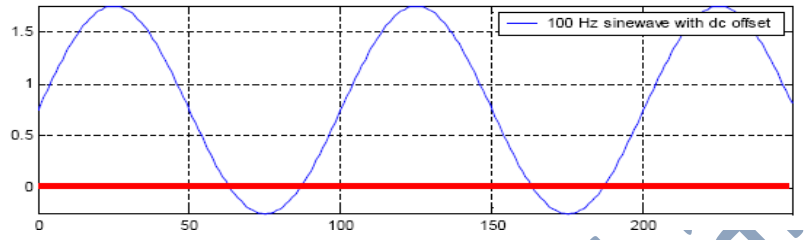
zero crossing => successive samples have different algebraic signs

ZC Example



ZC=9

Offset=0.75



ZC=8

ZC Definition

$$Z_n = \sum_{m=-\infty}^{\infty} |\,\mathrm{sgn}[x(m)] - \mathrm{sgn}[x(m-1)]\,|\,w(n-m)$$

$$\mathrm{sgn}[x(n)] = 1 \qquad x(n) \geq 0$$

$$= -1 \qquad x(n) < 0$$