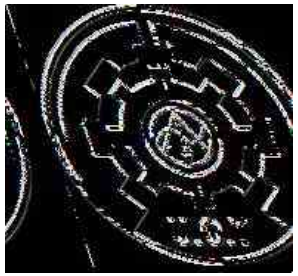
University of Technology
الجامعة التكنولوجية

Computer Science Department
قسم علوم الحاسوب

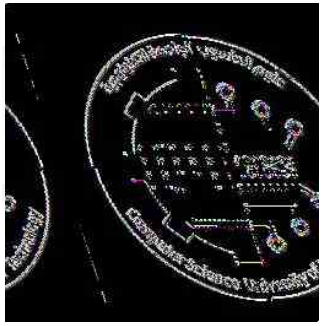
Speech Recognition
تمييز الكلام

Dr. Khitam A. Salman
د. ختام عبدالنبي سلمان

cs.uotechnology.edu.iq



Speech Recognition



2nd course lecture 1

Lecture Outlines

- ***Spoken language processing:***

  Understanding spoken language, Speech definition,

- ***problem areas in speech recognition system***
- **Word recognition:** speaker-dependent and speaker independent

Lecturer: Dr. Asia Ali

Speech Recognition

Speech is the most important manner of communication for Humans to exchange the information. Speech Recognition also called as voice recognition, which deals with analysis of the linguistic content of a speech signal and its conversion into a computer readable format.

The task of speech recognition is to convert speech into a sequence of words by a computer program.

Speech is composed of subsequent voice fragments. Similarly, in written expressions, these voice fragments are replaced with the symbols of the language, letters.

As the most natural communication modality for humans, the ultimate dream of speech recognition is to enable people to communicate more naturally and effectively. While the long-term objective requires deep integration with many NLP components discussed in the first course.

Many emerging applications have readily deployed with the core speech-recognition module. Some of these typical applications include voice dialing, call routing, command and control, and computer-aided language learning. Speech processing is a diverse field with many applications.

Most of these modern systems are typically based on statistic models such as hidden Markov models (HMMs). One reason why HMMs are popular is that their parameters can be estimated automatically from a large amount of data, and they are simple and computationally feasible.

Speech recognition in practice can either work on audio files or *convert a microphone waveform to a sequence of words.* Given the uncertainty at many levels of this problem (e.g., introduced by background noise, digitization noise, speaker's accent).

Spoken language processing:

Spoken language processing is beyond the scope of speech recognition. It is one of the most efficient, flexible and economical means of communication among humans**.**

Understanding spoken language:

Speech definition: is human vocal communication using language. Consists of a continuously varying sound wave which links speaker to listener. Sound requires a medium through which to travel and the most usual medium is air.

Types of Speech Recognition:

Speech recognition can be classified into speaker dependent or independent, isolated or continuous and can be for large vocabulary or small vocabulary.

In more detail, they are three styles of speech: isolated, connected and continuous.

A. **Isolated** speech means single words. Speech recognition systems can just handle words that are spoken separately. This is the most common speech recognition systems available today. The user must pause between each word or command spoken. The speech recognition circuit is set up to identify isolated words of 96-second lengths.
B. **Discontinuous**/connected speech means full sentences in which words are artificially separated by silence i.e. it is a halfway point between isolated word and continuous. Speech recognition Allows users to speak multiple words. The HM2007 can be set up to identify words or phrases 1.92 seconds in length. This reduces the word recognition vocabulary number to 20.
C. **Continuous** speech means naturally spoken sentences. It is extremely difficult for a recognizer to shift through the text as the word tend to merge together. For instance, "Hi, how are you doing?" sounds like "Hi,.howyadoin" Continuous speech recognition systems are on the market and are under continual development.

Difficulties of Speech Recognition / *problem areas in speech recognition system*

The main goal of a speech recognition system is to substitute for a human listener, although it is very difficult for an artificial system to achieve the flexibility offered by human ear and human brain. Thus, speech recognition systems need to have some constraints. For instance, number of words is a constraint for a word-based recognitions system. In order to increase the performance of the recognition, the process is dealt with in parts, and researches are concentrated on those parts. This

approach of splitting the process into parts provide better performance achievement for each of the parts, thus resulting in increased overall performance.

However, the following problems is still existed.

Speech is highly variable even for the same speaker

-Variable pitch contour (e.g. surprise, anger)

-Variable speed

-Effect of having a blocked nose.

Different speakers pronounce words differently

-Accents give gross changes

-Smaller changes within a single accent

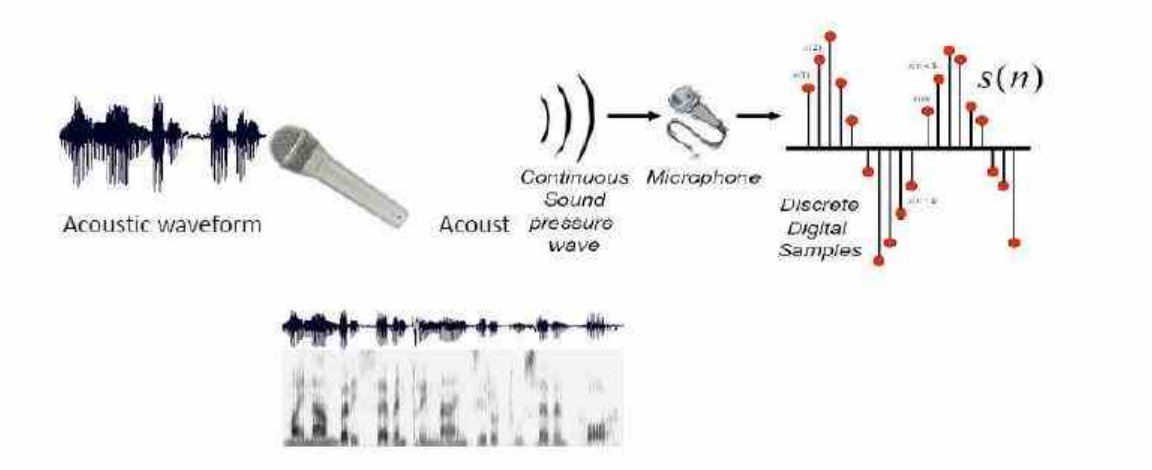
Speech sounds vary according to context

“handbag” → “hanbag” → “hambag”

Natural speech includes extraneous sounds Coughs, “umm”, “err”, false starts.

How might computers do it?

- Digitization.
- Acoustic analysis of the speech signal.
- Linguistic interpretation.



Structure of a standard speech recognition system

1-Raw speech:

Speech is typically sampled at a high frequency, e.g., 16 KHz over a microphone.

2-Signal analysis

Raw speech should be transformed and compressed, in order to simplify subsequent processing. We say that a speech representation is robust if it is (almost) identical each time the same word is spoken. Different utterances of the same word *must give* similar feature vectors. Different words must give distinctly different feature vectors.

3- Pattern-matching approach

The *pattern-matching approach* involves two essential steps:

1-pattern training

2-Pattern Comparison.

The essential feature of this approach is that it uses a well formulated mathematical framework and establishes consistent speech-pattern representations, for reliable pattern comparison, from a set of labeled training samples via a formal training algorithm.

In the pattern-comparison stage of the approach, a direct comparison is made between the unknown speeches (the speech to be recognized) with each possible pattern learned in the training stage in order to determine the identity of the unknown according to the goodness of **match of the patterns**:

1-Matching method: Dynamic programming DP

2-statistical model: Hidden Markov Model (HMM)

3-Neural network.

Word recognition: speaker-dependent and speaker independent

Word recognition

There have been two major drives at speech recognition;

Speaker dependent isolated word recognition (the simplest one). Reference patterns are constructed for a single speaker. In order for a system to recognize the speech of different speakers, the reference patterns must be updated for new speakers Speaker-dependent recognition helps to overcome such problems as regional, accent, the sex of the speaker, etc.
Hence, Speaker-dependent solutions are found in specialised use cases where there a limited number of words that need to be recognized with high accuracy.

Speaker-dependent isolated-word recognition involves training the computer to recognize words by getting the speaker repeatedly to say certain words. From the results of the initial training the computer is able to formulate an average template for individual words, which are then used for reference. Obviously, the more training a machine receives, the better it becomes at choosing the right word(s).

Each word must be spoken separately, as opposed to continuously, which is the case in normal speech. The problem here is that the user may not always be prepared to help the machine in this way, since the process can be very time-consuming.

Speaker-independent isolated word recognition. The system is able to recognize the speech of any speaker. Unfortunately, it is much more difficult to develop such a system than a speaker-dependent one.

Speaker-independent recognition involves

- converting the spoken word into an electrical signal of some sort via a microphone.
- the signal is then processed further; to obtain a set of identifying features
- Then are compared with those held in the computer's vocabulary: The vocabulary will consist of asset of reference templates, which have been chosen to represent the average speaker, or speakers with similar accents.

speaker–independent software ideal for most IVR (**Interactive Voice Response**) systems, and any application where a large number of people will be using the same system. It is more often found in telephone applications.

In both cases (*speaker-dependent and speaker-independent recognition*) some degree of pattern matching is necessary in order to recognize a word. The system compares the incoming signal with a stored template, there by generating some sort of score on the basis of the degree of similarity. The template with the highest score is then chosen.

Another classification can be made according to the size of the recognition unit chosen, as mentioned before. Speech recognition systems can be separated into two groups;

- In a **Word Based Speech Recognition**, the smallest recognition unit is a **word**. Recognition accuracy of the first one is very high because the system is free from negative side effects of co-articulation. However, for continuous speech recognition, transition effects between words again cause problems. Moreover, for a word-based recognition system, processing time and memory requirements are very high because there are many words in a language, which are the bases of the reference patterns.

- Phoneme Based Speech Recognition systems are the ones that use phonemes as the recognition units. In a phoneme-based system, while recognition accuracy
decreases, it is possible to apply error-correction using the ability to produce fast results with very few phoneme numbers.
There can be several speech recognition systems that make use of sub-word units like diphone-based, triphone-based, and syllable-based

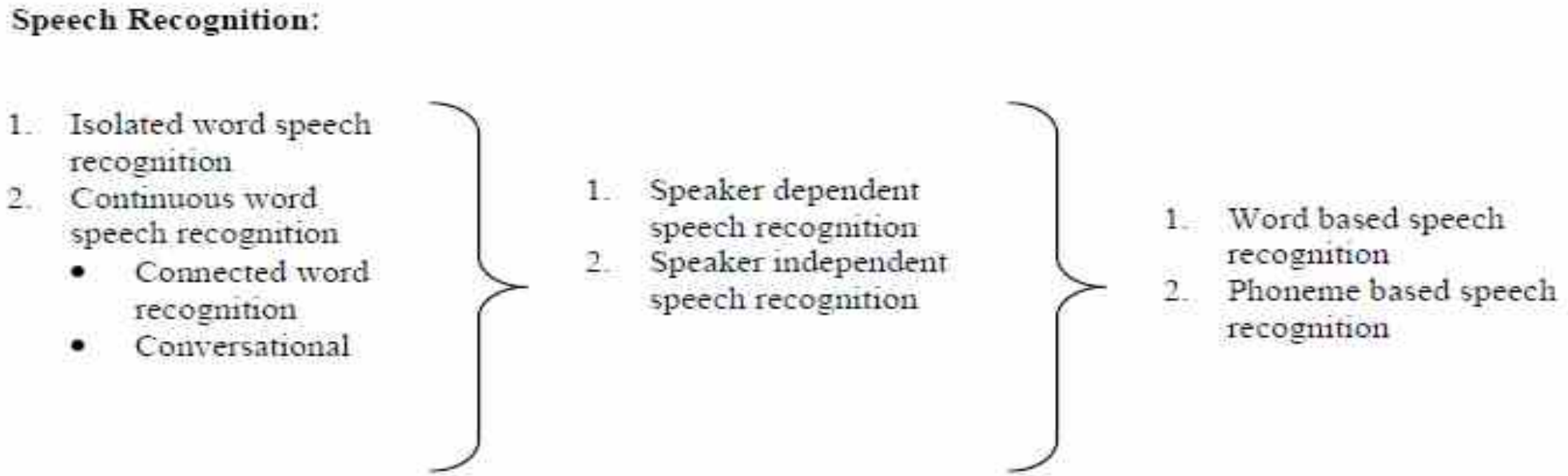
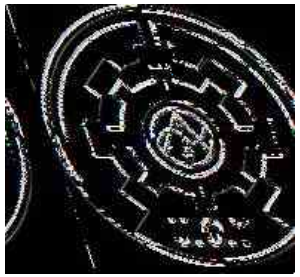
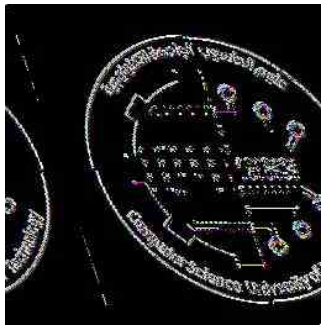


Figure 1- Classification of Speech Recognition Systems



Speech Recognition



2nd course lecture 2

Lecture Outlines

- Speech Recognition tasks
- Types of speech recogniser
- Variability in speech recognition
- Basic step of speech recognition

Lecturer: Dr. Asia Ali

➢ Speech Recognition tasks

Speech to text: transforming the recorded audio/ human speech in to a sequence of words an accurate written transcript, just the words no meaning is required. Hence, transcriptions may be in words, phonemes, syllables, or other units. We need to deal with acoustic ambiguity (recognize the speech).

Speaker Dualization (who spoke when)

Is this a woman or a man?

Segmenting a dialogue or multiparty conversation

- Speech recognition: what did they say?
- Paralinguistic aspects: how did they say it? (timing, intonation, voice quality)
- Speech understanding: what does it mean?

– Possibility that speaker is not among the speakers known to the system.

➢ Types of speech recogniser

Speech recognition is normally categorised by a few key descriptive phrases:

- **Automatic speech recognition (ASR)** describes a system that can recognise speech without additional user input.
- **Continuous speech recognition** describes a speech recognition system that can recognise continuous sentences of speech. In theory this would not require a user to pause when speaking and would include dictation and transcription systems. The alternative is a **discrete word** recognition system, used primarily for handling vocal commands, that recognises single words delimited by pauses.
- **Natural language processing (NLP)**, whilst not strictly limited to speech, describes the computational methods needed for a computer to understand the meaning of what is being said, rather than simply knowing what words have been said.

➢ Variability in speech recognition

Several sources of variation

- Size Number of word types in vocabulary, perplexity
- Speaker Tuned for a particular speaker, or speaker-independent? Adaptation to speaker characteristics
- Acoustic environment Noise, competing speakers, channel conditions (microphone, phone line, room acoustics)
- Style Continuously spoken or isolated? Planned monologue or spontaneous conversation?
- Accent/dialect Recognize the speech of all speakers who speak a particular language
- Language spoken There are many languages beyond English, Mandar in Chinese, Spanish, . . .etc.

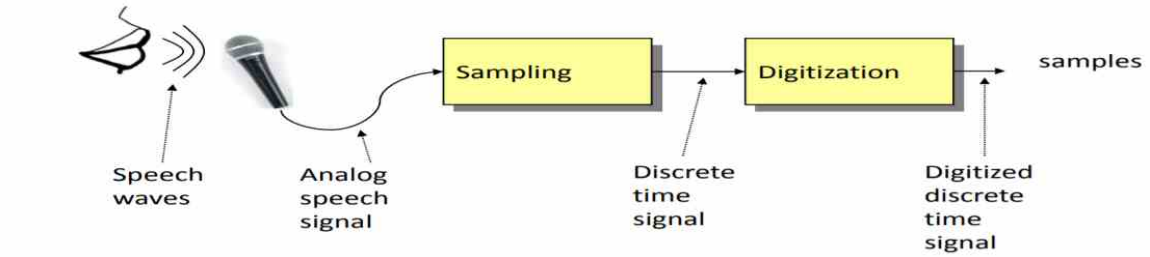
➢ Basic step of speech recognition*:*

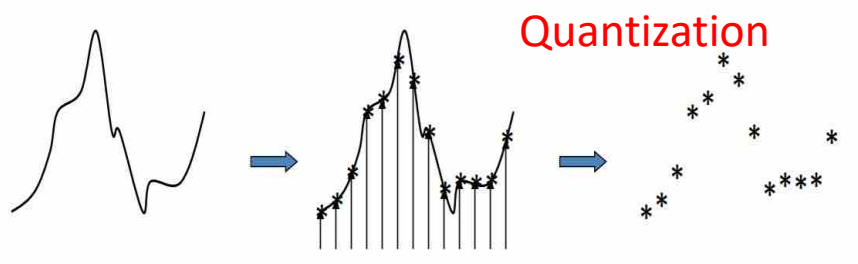
1. Sampling
2. Signal detection
3. Speech spectra
4. Pitch contour evaluation
5. Segmentation
6. Word recognition
7. Responding to the message

1. Sampling

The signal from the microphone goes through a pre-amp. The gain of which can be adjusted.

- **The output of the pre-amp is a continuous-time electrical signal-** Usually a voltage signal.
- **The signal is digitized by an analog to digital converter** – The signal value is send at regular, periodic intervals of time
- **Sampling** – The value at each instant is quantized to one of a number of fixed values
- **Quantization –**





- Sampling is the process of capturing snapshots of the signal at discrete instants of time

– For speech recognition, these samples will be uniformly spaced in time

• Requirement: The sequence of numbers must be sufficient to reconstruct the original signal perfectly.

– To retain all the information in the signal

Having obtained the digital representation of the incoming signal, various parametric features can then be obtained. These parametric features then form the basis for the segmentation of the speech signal. Further, to avoid distortion of the incoming signal by background noise, it is desirable to have the microphone as close as possible to the speaker, if it is not possible to operate in a noise-free environment.

2. Signal Detection

The signal processor remains in waiting state when no voice signals high enough to be detected are coming from the receiver (*microphone*).

Variation in this silent environment is automatically detected by the program in execution and the speech recognition process is activated.

A similar principle applies when the end of an incoming signals Detected, that is, when there is a variation from a signals silence, and the signal is not of sufficient strength to energize the filters; the speech recognition program recognizes this as the end of the utterance.

3. Speech spectra

It is the average sound spectrum for the human voice. The analysis of the speech signal is always the foundation of related processing techniques. Since speech signal is time-varying, the analysis should be a time-frequency analysis. Short-time FT (STFT) is applied in the spectral analysis for speech.

Sound spectra graph utilizes a special frequency analyser to produce a three-dimensional plot of the variation in the speech energy spectrum with time.

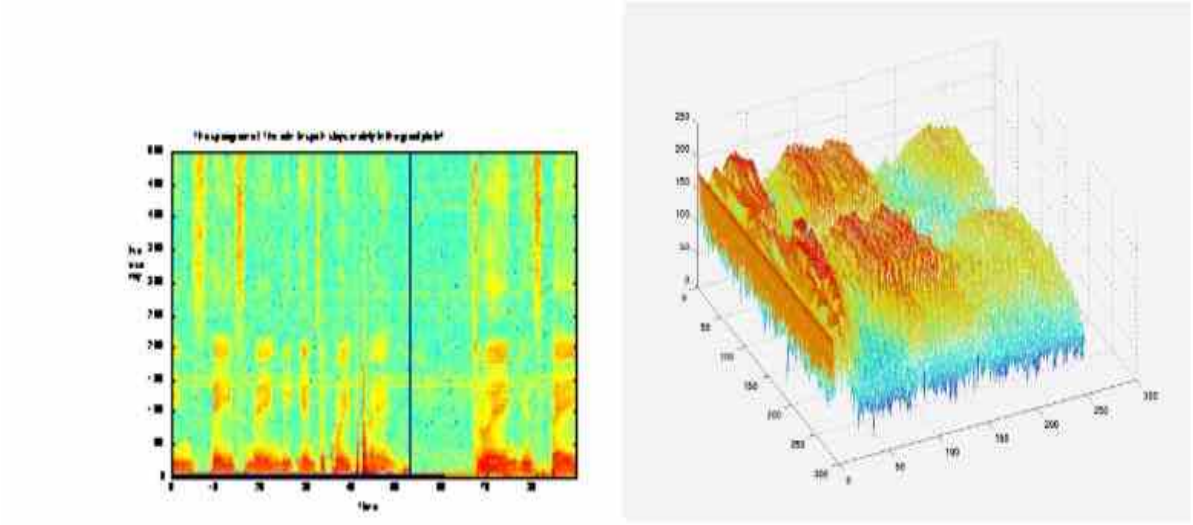


Figure 1-The spectra gram style. - The 3-D mesh style

Voiced/Unvoiced Spectrum

Speech can be generally divided into voiced and unvoiced. We studied both the spectrum of typical voiced and unvoiced block. The spectrum in Fig.1. shows that for a voiced speech, the time series have obvious periodic. The spectrum of voiced speech is featured as some fine spectrum with formant envelope. The fine peaks mean the pitch period and the formants reflect the vocal tract feature. While for the unvoiced case, the signal looks much like a white noise.

4. Pitch Contour Evaluation

Pitch is the relative highness or lowness of a tone as perceived by the ear, which depends on the number of vibrations per second produced by the vocal cords. Pitch is the main acoustic correlate of tone and intonation. In a quiet environment, pitch conveys important linguistic information for the perception of speech information, including phonemes and words.

Tone, in linguistics, a variation in the pitch of the voice while speaking.

The contours of the fundamental frequency (as it rises and falls) can also be used as an indication of major syntactic boundaries. Additionally, stress patterns, rhythm and intonation carry some clues as to the Phonetic identify of some speech sounds.

5. Segmentation

Segmentation: Segmentation is the process of decomposing the speech signal into a set of basic phonetic units. Speech signal can be segmented into **words, sub words, syllables and phonemes**. Speech segmentation was done using wavelet, fuzzy methods, artificial neural network, hidden markov model.

Segmentation is used to segments continuous speech into uniquely identifiable or phonemes, syllables, words or sub words and processes them to generate distinguishable features.

Segmentation is an important role in speech recognition:

- reduce memory size and computational complexity for large vocabulary systems.
- Segmentation is used to detect the proper start and end point of speech events.

There Are Two Kinds of Segmentation:

- Phonemic segmentation
- syllable like unit segmentation.

Phonemic segmentation segments speech sequence into small phonemes and aided segmentation segments speech into small syllables.

Allophone

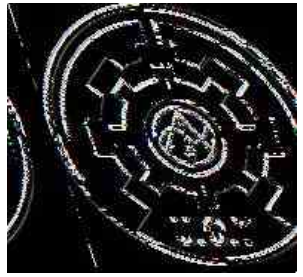
any of the speech sounds that represent a single phoneme, such as the aspirated k in kit and the unaspirated k in skit, which are allophones of the phoneme k. Certain allophones can provide word or syllable boundary information which can be, very useful in recognition systems. For example, some allophonic variations occur only at the end or beginning of words/syllables.

Phoneme

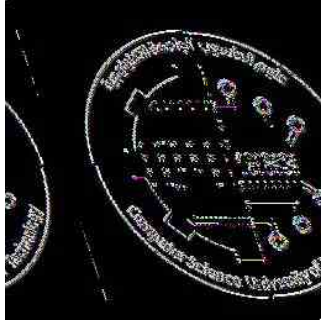
The phoneme represents the smallest number of distinctive phonological classes for recognition and is substantially less than the number of allophones, syllables, words, etc. (there are approximately 43 phonemes in the English language). Although the number of phonemes is small, their automatic recognition by a computer system is still a major problem, since there are no a caustically well-defined patterns or templates for phonemes.

Diphone

term diphone is used to represent a vowel-consonant sequence, such that the segment is taken from the centre of the vowel to the center of the consonant.



Speech Recognition



2nd course lecture 3

Lecture Outlines

- Word Recognition
- Responding To The Message
- Automatic Speech Recognition (ASR)
- Feature Extraction

Lecturer: Dr. Asia Ali

Word Recognition

In general, Word recognition is a process by which students learn to identify words and word parts. It begins with an understanding that letters symbolize the sounds in words and progresses to the ability to understand complex word parts and syllabication principles.

- If segmentation takes place at a level lower than a word, some means of getting from the lower level to that of stored words has to be included in the system. There is a certain degree of uncertainty when trying to assign words to segmented data, and in many cases it is possible to assign more than one word to the data.

When this occurs, an algorithm has to be included in the system which can generate the most likely word.

The algorithm may include some form of syntactic analysis in order to:

- weed out ungrammatical sequences.
- gain some knowledge about the most likely words in a particular context.

This can, however, make the recognition system very task specific.

- Syntactical analysis can also be used to restrict the recognition of the next word on the basis of *previously stored words*. This is desirable, because with large lexicons the task of searching for the best match can become very expensive computationally.

However, due to the vagaries of the English language, this syntactical approach has certain limitations, including the difficulty in distinguishing between well-formed and poorly formed sentences.

A statistical approach can be adopted at all levels of decision making whereas core can be assigned to each of the alternatives on the basis of past history; the alternatives with the highest score being the one selected for further processing.

There are many ways of obtaining the highest score.

- ❑ The breadth first search computes the score at each alternative and selects the route with the highest score.
- ❑ The depth first search selects the highest score at the initial level and then pursues this initial choice in subsequent levels, in a depth first manner. The problem with the depth first technique is that the system **is** committed to the consequences of the first choice.
- ❑ There are also searching techniques which are a hybrid of breadth first and depth first techniques.

Responding To The Message

Assuming that all the words and sentences have been correctly identified, the computer must then be able to respond in the appropriate manner. The response can be in the form of an execution of an operating system command, the display of a string (message) on the screen, the lexical display of the words actually spoken, etc.

➢ Automatic Speech Recognition (ASR)

The goal of an ASR system is to accurately and efficiently convert a speech signal into a text message transcription of the spoken words, independent of the device used to record the speech i.e. the microphone, the speaker's accent, or the acoustic environment in which the speaker is located (e.g., quiet office, noisy room, outdoors).

Many core technologies, such as Gaussian mixture models (GMMs), hidden Markov models (HMMs) and various adaptation techniques have been developed along the way.

The increased demands on ASR in mobile devices and the success of new speech applications in the mobile world such as voice search (VS), short message dictation (SMD), and virtual speech assistants (e.g., Apple's Siri, Google Now, and Microsoft's Cortana). The most popular applications in this category include voice search, personal digital assistant, gaming, and living room interaction systems.

Basic Architecture of ASR Systems

A practical ASR system, nowadays, needs to deal with huge (millions) vocabulary, free-style conversation, noisy far field spontaneous speech and mixed languages.

The typical architecture of an ASR system is illustrated in Fig 1. As indicated in the figure, the ASR system has four main components: ***Signal processing and feature extraction, acoustic model (AM), language model (LM), and hypothesis search.***

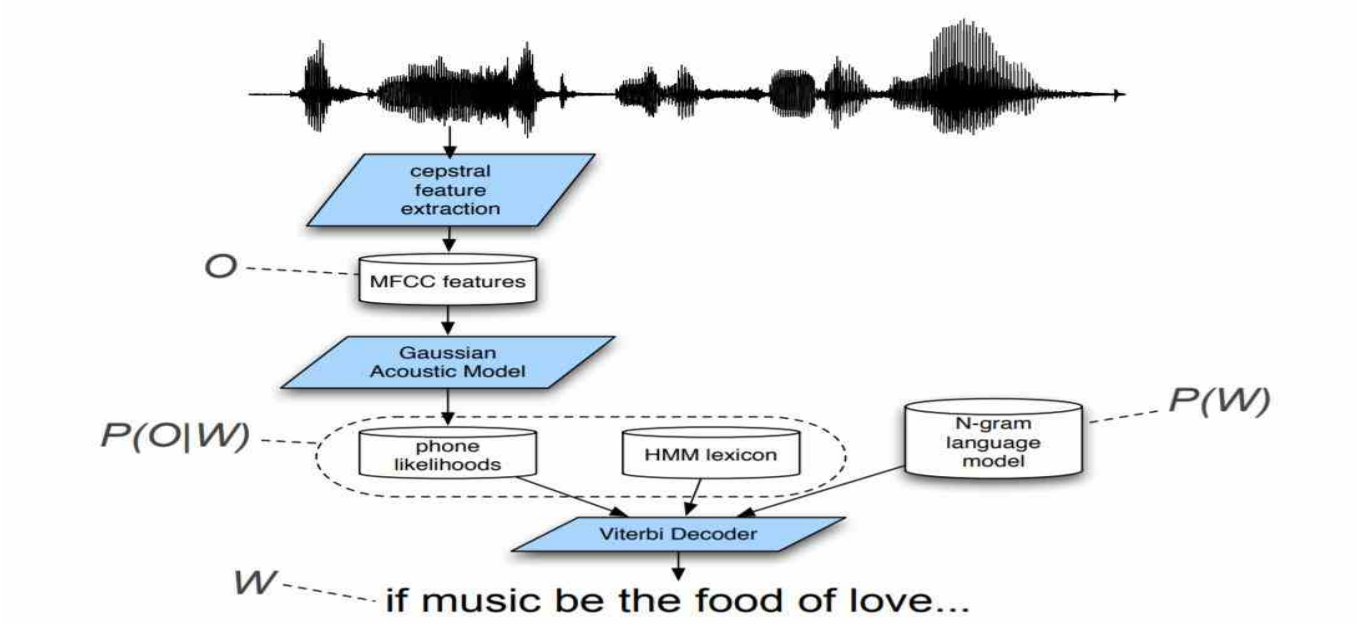


Figure 1- Speech recognition architecture

➢ Feature Extraction

In feature extraction, there is two types of signals are used.

The extracted features must have specific characteristics: Easily measurable, occur naturally and frequently in speech. Not change over time. Vary as much among speakers, consistent for each speaker. Not affected by: speaker health, background noise.

After features are extracted, a simple thresholding method is used to detect the word boundaries. The segmentation is used to divide the entire speech sequence into a sequence of words or sub words. Among these methods, spectral centroid has high segmentation accuracy.

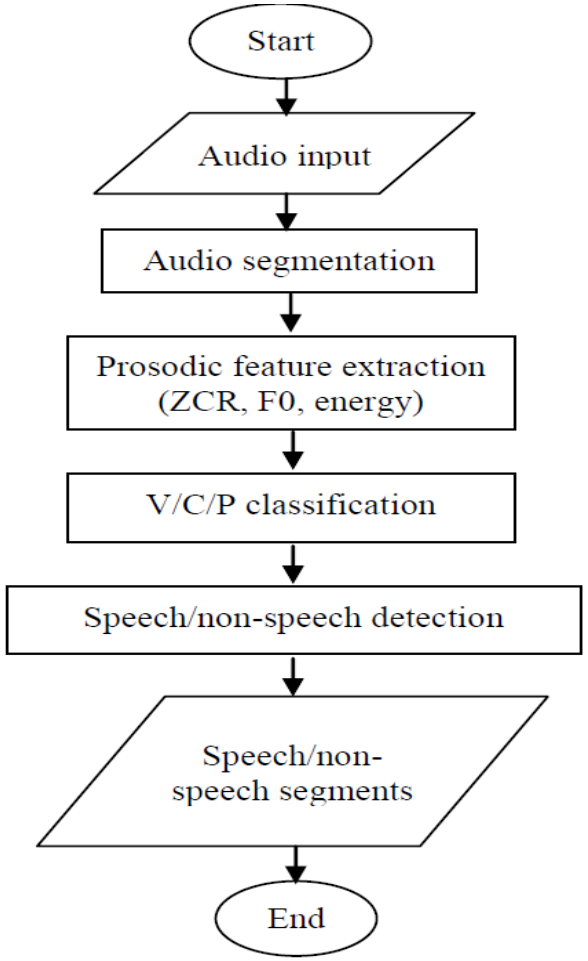


Fig. 2. Speech/Non-Speech detection flowchart

➢ Time Domain Features

1- Zero Crossing Rate

Zero crossing rates (ZCR) is measured based on the number of times the audio signal crosses the zero amplitude line by transition from a positive to negative or vice versa. Frames that have high number of times are categorized as speech segments and frames with low number of times are categorized as non-speech segments.

A ZCR threshold value is calculated to determine the speech/non-speech segments. The zero crossing value for the k-th segment is computed using Eq.(1), where can be three possible value that is +1, 0, -1 depending on whether the sample is positive, zero or negative.

$$Z_k = \sum_{n=1}^{N-1} |sgn[x_i(n)] - sgn[x_i(n-1)]| \qquad (1)$$

where

$$sgn[x_i(n)] = \begin{cases} +1, x_i(n) \geq 0 \\ -1, x_i(n) < 0 \end{cases} \qquad (2)$$

2- Fundamental Frequency

Fundamental frequency (F0), measured in Hertz is defined as the lowest frequency of a periodic waveform. A period of the waveform is the shortest possible time after which the waveform repeats itself. This single period is the smallest repeating unit and it will describe the signal completely. The equation of F0 is defines as in Eq. (3);

$$F0 = \frac{1}{T} \qquad (3)$$

where $F0$ is the fundamental frequency and T s the fundamental period. Frames that have very low value of F0 are categorized non-speech segments and frames with high F0 are categorized as speech segment. We extract F0 in MATLAB

3- Short Time Energy

Energy is very much related to the amplitude. It is a way of representing the amplitude changes in speech signal. The segment of a speech signal such as non-speech segment have much lower amplitude than the speech segment, resulting non-speech segment to have lower energy. Therefore, the energy measure can be used to discriminate between speech and non-speech segments with selected set of threshold.

Energy (E_k) for *k*-th segment is defined in Eq. (4), where g(t) is the amplitude for t-th frame and *N* is the number of frames.

$$E_k = \sum_{n=1}^{N-1} |g(t)|^2 \qquad (4)$$

Frequency Domain

Spectral Flux

Spectral flux refers to a measure of how quickly the power spectrum of a signal is changing, calculated by comparing the power spectrum for one frame against the power spectrum from the previous frame. The spectral flux can be used to determine the timbre of an audio signal.

The Spectral Flux is given by

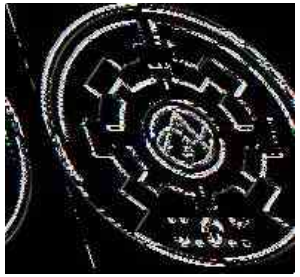
$$SF_i = \sum_{k=1}^{N/2} \left(|X_i(k)| - |X_i(k-1)| \right)^2 \quad \ldots\ldots (8).$$

Here, X i (k) is the DFT coefficients of i-th frame.

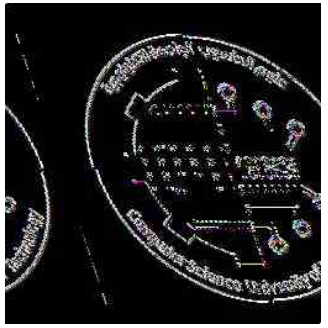
The signal processing and feature extraction component do the following

1. Takes as input the audio signal.
2. Enhances the speech by removing noises and channel distortions
3. Converts the signal from time-domain to frequency-domain, and extracts salient feature vectors that are suitable for the following acoustic models.

The three prosodic features F0, energy and ZCR have their own advantages and disadvantages in detecting the speech/non-speech segments. F0 and energy can be used to differentiate speech segments and non-speech segments. On the other hand, ZCR is more useful in classifying the speech segments into vowel/consonant/pause (V/C/P). V/C/P classification is very important for a more accurate detection of speech/non-speech segments and for calculation of rate of speech (ROS) in sentence boundary detection.



Speech Recognition



2nd course lecture 4

Lecture Outlines:

- Acoustic Model
- Language Model
- The Noisy Channel Model
- The Decoding Phase
- Speech Processing Task

Lecturer: Dr. Asia Ali

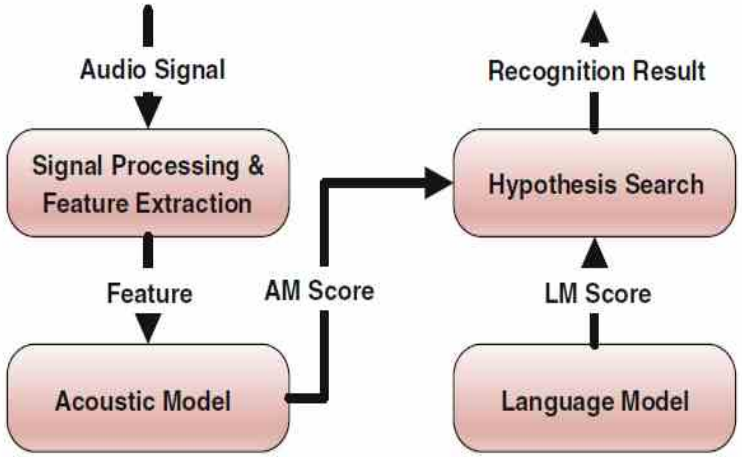
The Acoustic Model

1. Integrates knowledge about acoustics and phonetics
2. Takes as input the features generated from the feature extraction component, and generates an AM score for the variable-length feature sequence.

The Language Model

1. Estimates the probability of a hypothesized word sequence, or LM score, by learning the correlation between words from a (typically text) training corpora. The LM score often can be estimated more accurately if the prior knowledge about the domain or task is known.

The hypothesis search component combines AM and LM scores given the feature vector sequence and the hypothesized word sequence, and outputs the word sequence with the highest score as the recognition result.



The Noisy Channel Model

The task of speech recognition is to take as input an acoustic waveform and produce as output a string of words.

HMM based speech recognition systems view this task using the metaphor of the noisy channel.

The intuition of the **noisy channel model (Fig.1 is to treat the acoustic waveform as an** "noisy" version of the string of words, i.e. A version that has been passed through a noisy communications channel.

This channel introduces "noise" which makes it hard to recognize the "true" string of words.

Our goal is then to build a model of the channel so that we can figure out how it modified this "true" sentence and hence recover it.

The insight of the noisy channel model is that if we know how the channel distorts the source, we could find the correct source sentence for a waveform by taking every possible sentence in the language, running each sentence through our noisy channel model, and seeing if it matches the output.

We then select the best matching source sentence as our desired source sentence.

In summary:

- Search through space of all possible sentences.
- Pick the one that is most probable given the waveform.
- We search through a huge space of potential "source" sentences and choose the one which has the highest probability of generating the "noisy" sentence.
- We need models of the prior probability of a source sentence, the probability of words being realized as certain strings of phones (HMM lexicons), and the probability of phones being realized as acoustic or spectral features (Gaussian Mixture Models).

Implementing the noisy channel model as we have expressed it in Fig.1 requires solutions to two problems.

- **First**, in order to pick the sentence that best matches the noisy input we will need a complete metric for a “best match”. Because speech is so variable, an acoustic input sentence will never exactly match any model we have for this sentence. Probability will be used as metric.

Second since these to fall English sentences is huge, an efficient algorithm needed that will not search through all possible sentences, but only ones that have a good chance of matching the input. This is the decoding or search problem**.**

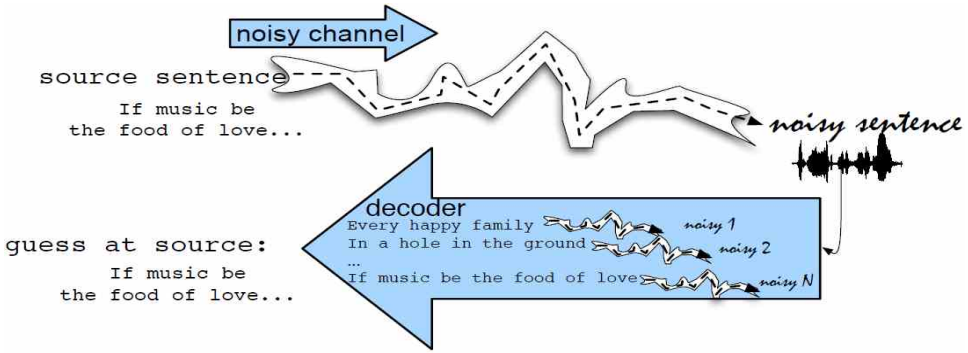
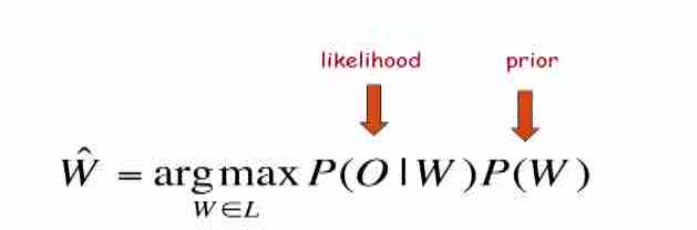


Figure 1- The noisy channel model.

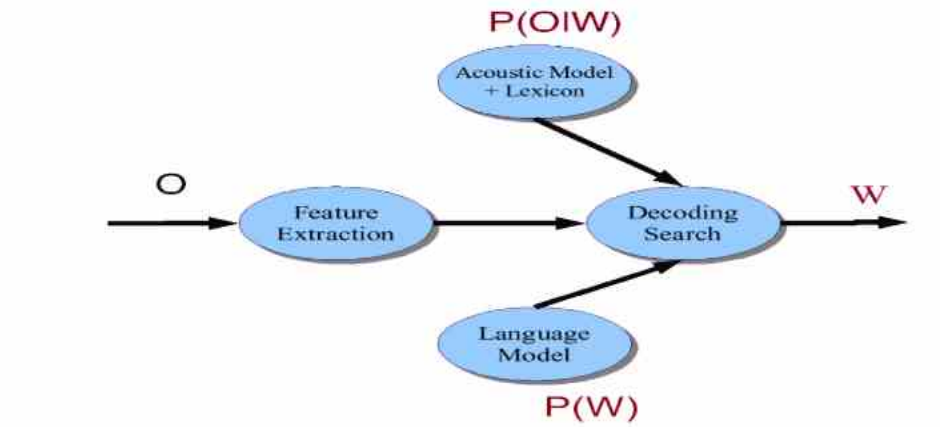
• Many problems involve predicting a complex label Y from data X

- in automatic speech recognition, X is acoustic waveform, Y is transcript
- in machine translation, X is source language text, Y is target language translation
- in spelling correction, X is a source text with spelling mistakes, and Y is a target text without spelling mistakes
- in automatic summarization, X is a document, Y is a summary of that document.

In Language Model



Speech Architecture meets Noisy channel



The above figure shows further details of the operationalization in, which shows the components of an HMM speech recognizer as it processes a single utterance.

The figure shows the recognition process in three stages. In the **feature extraction or signal processing stage, the acoustic wave form is sampled** into **frames (usually of 10, 15, or 20 milliseconds) which are transformed** into **spectral features. Each time window is thus represented by a vector of** around 39 features representing this spectral information as well as information about energy and spectral change.

In the **acoustic modelling or phone recognition stage, we compute the** likelihood of the observed spectral feature vectors given linguistic units (words, phones, sub parts of phones). For example, we use Gaussian Mixture Model (GMM) classifiers to compute for each HMM state *q, corresponding* to a phone or sub phone, the likelihood of a given feature vector given this phone $p(o|q)$.

A(simplified) way of thinking of the output of this stage is as a sequence of probability vectors, one for each time frame, each vector at each time frame containing the likelihoods that each phone or sub-phone unit generated the acoustic feature vector observation at that time.

Finally, in **The Decoding Phase, we take the acoustic model (AM),** which consists of this sequence of acoustic likelihoods, plus an HMM dictionary of word pronunciations, combined with the language model (LM) (generally an *N-gram grammar),and output the most likely sequence of* words. An HMM dictionary, is a list of word pronunciations, each pronunciation represented by a string of phones.

Each word can then be thought of as an HMM, where the phones (or sometime sub phones) are states in the HMM, and the Gaussian likelihood estimate or supply the HMM output likelihood function for each state. Most ASR systems use the Viterbi algorithm for decoding, speeding up the decoding with wide variety of sophisticated augmentations such as pruning, fast-match, and tree-structured lexicons.

Example: ASR Architecture: Five easy pieces: ASR Noisy Channel architecture

. **Feature Extraction:**. "MFCC" features
. **Acoustic Model:**. Gaussians for computing p(o|q)
. **Lexicon/Pronunciation Model:**. HMM: what phones can follow each other
. **Language Model:**. N-grams for computing p(wi|wi-1)
. **Decoder:**. Viterbi algorithm: dynamic programming for combining all these to get word sequence from speech

Speech Processing Task

The two main issues to deal with by the AM component are the variable-length feature vectors and variability in the audio signals.

The variable length feature problem is often addressed by techniques such as dynamic time warping (DTW) and hidden Markov model (HMM).

As illustrated in Fig. 2, a practical ASR system, nowadays, needs to deal with huge (millions) vocabulary, free-style conversation, noisy far field spontaneous speech and mixed languages.

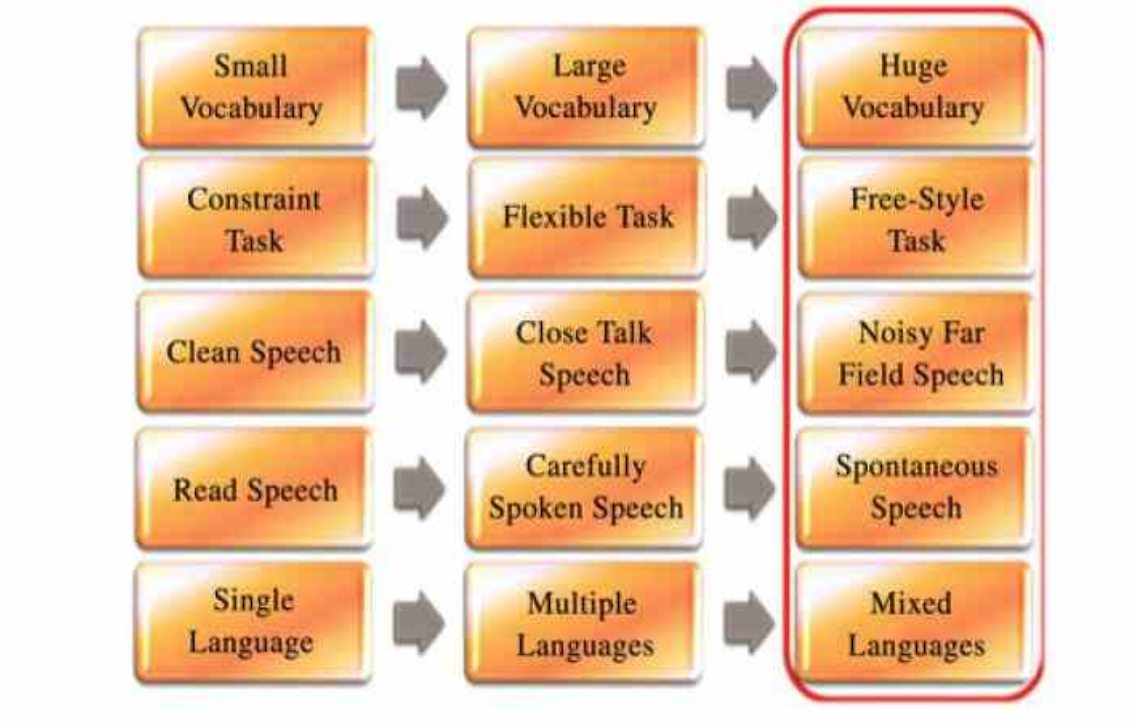
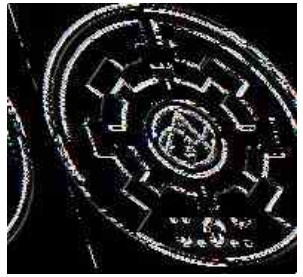
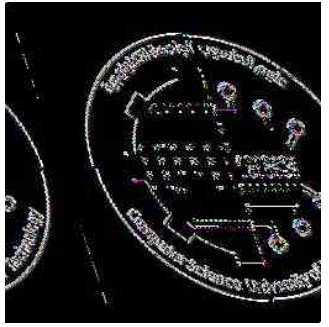


Figure 2- The ASR problems we work on today (right column) are much more difficult than what we have worked on in the past due to the demand from the real-world applications.



Speech Recognition



2nd course lecture 5

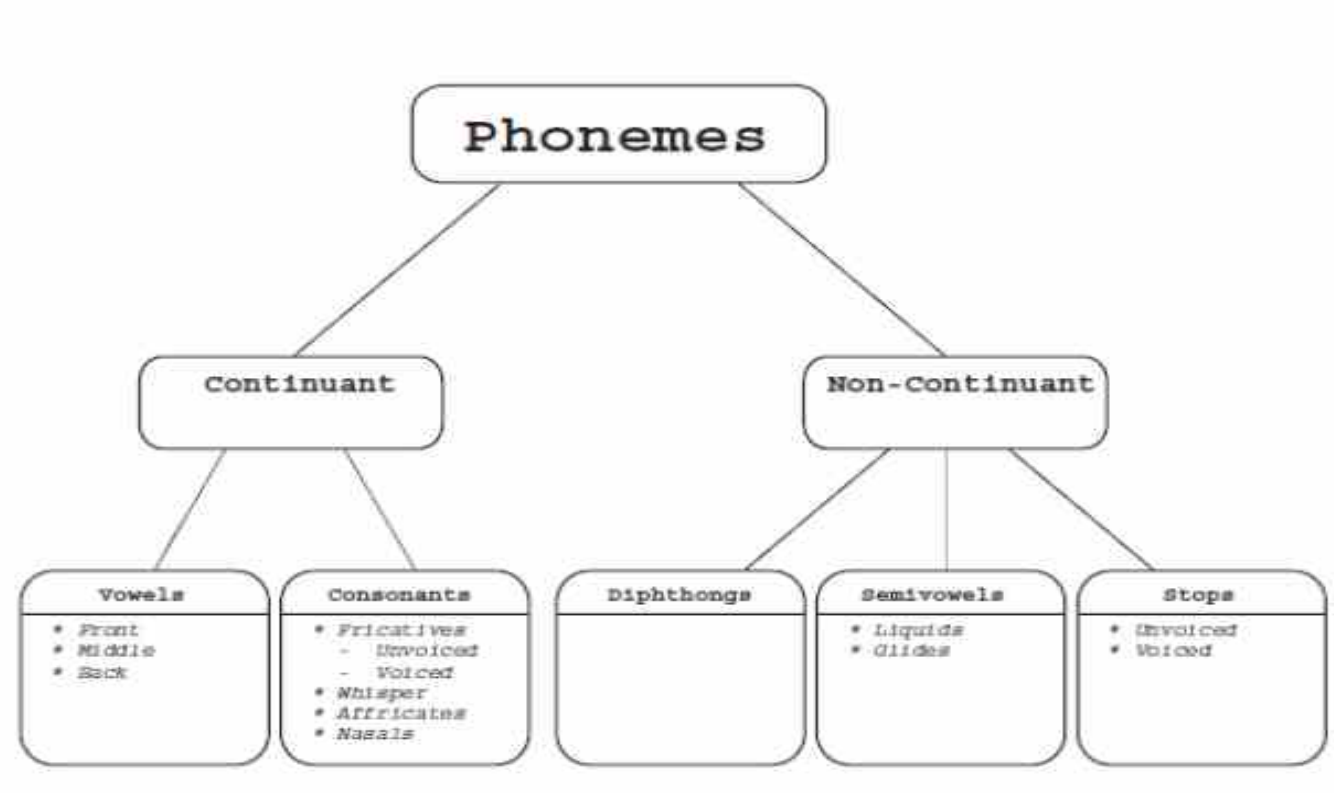
Lecture Outlines:

- **ASR Summary**
- **HMM**
- **Speech Synthesis**

Lecturer: Dr. Asia Ali

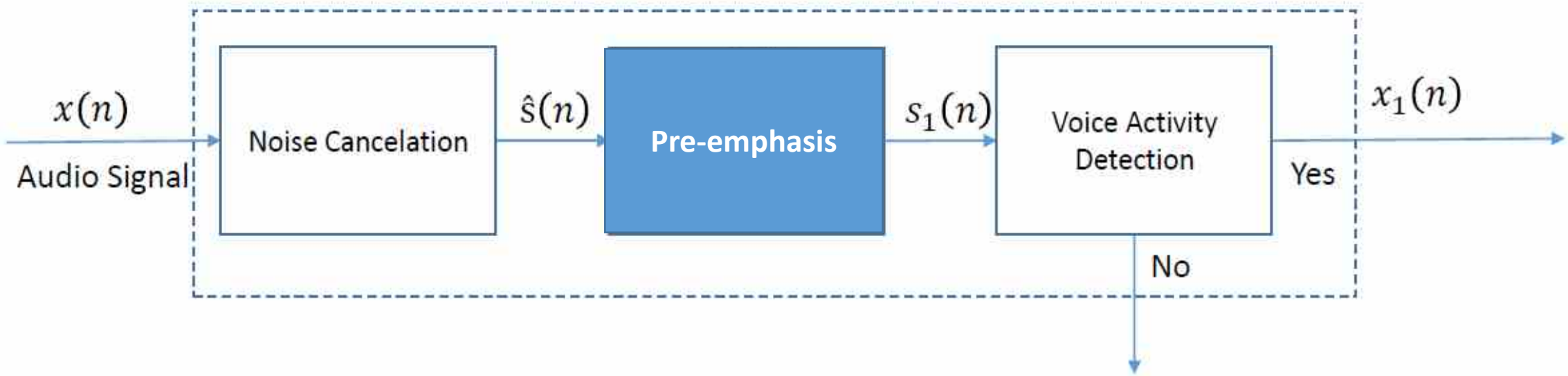
Speech signal

- The voice band is 0-4000 Hz
- Phoneme is the smallest unit of a phonetic language



❖ ASR Model

1- ASR Model – Pre-processing



It is assumed that the initial part of the signal (usually 200 ms) is noise and silence.

$$VAD(m) = \begin{cases} 1, & W_{s_1}(m) \geq t_w \\ 0, & W_{s_1}(m) < t_w \end{cases}$$

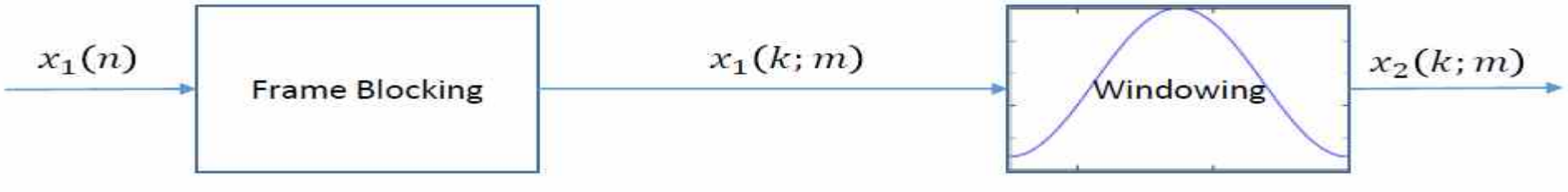
, where t_w is equal with W_{s_1} computed for the first 200 ms. The W_{s_1} usually depends on signal energy, power, zero crossing rate

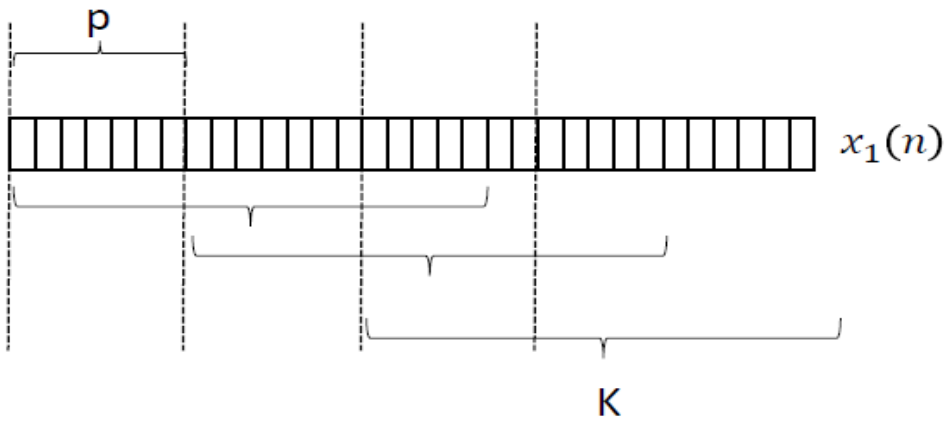
2- ASR Model – Features Extraction.

Frame Blocking and Windowing

Frame Blocking : Each frame is K samples long and overlaps the previous one with P samples

Windowing: In order to remove discontinuities a Hamming window is applied





In order to remove discontinuities a Hamming window is applied

3- ASR Model – Detection Engine

- Pattern matching
- Hidden Markov Models
- Neural Networks

Hidden Markov Models (HMM)

HMM is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (i.e. hidden) states.

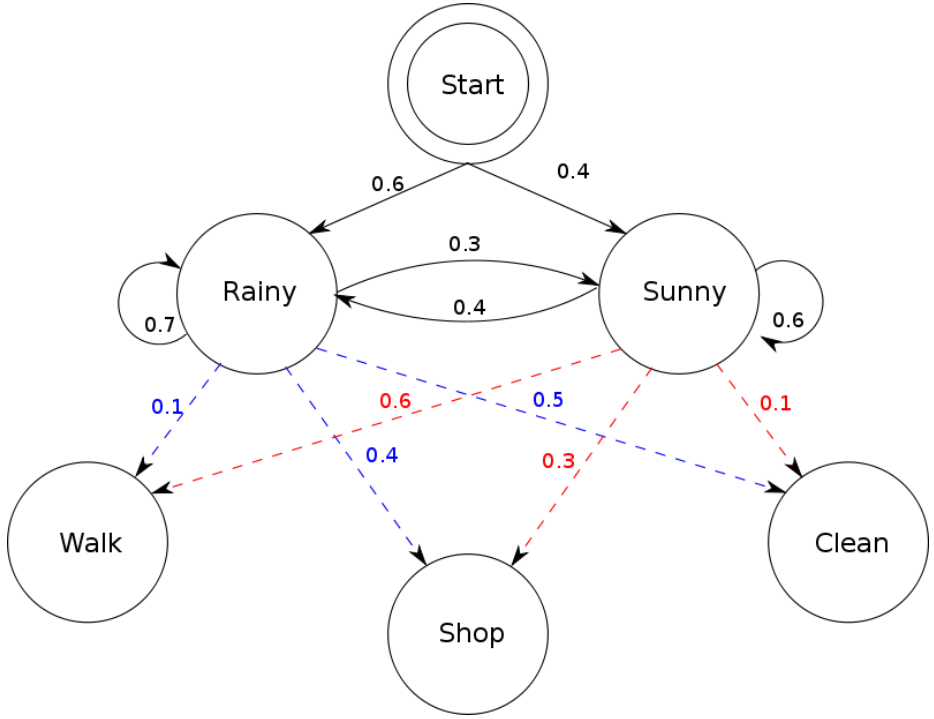
Hidden Markov models are especially known for application such as reinforcement learning and temporal pattern recognition such as speech recognition, part-of-speech tagging,..etc.

HMM is a two layer model first Hidden state and the second is Observation state).

Terminology in HMM

The term hidden refers to the first order Markov process behind the observation. Observation refers to the data we know and can observe. Markov process is shown by the interaction between “Rainy” and “Sunny” in the below diagram and each of these are **Hidden States**.

T = length of the observation sequence
N = number of states in the model
M = number of observation symbols
Q = $\{q_0, q_1, \ldots, q_{N-1}\}$ = distinct states of the Markov process
V = $\{0, 1, \ldots, M-1\}$ = set of possible observations
A = state transition probabilities
B = observation probability matrix
π = initial state distribution
$\mathcal{O}$ = $(\mathcal{O}_0, \mathcal{O}_1, \ldots, \mathcal{O}_{T-1})$ = observation sequence.



T = don't have any observation yet, N = 2, M = 3, Q = {"Rainy", "Sunny"}, V = {"Walk", "Shop", "Clean"}

Observations

are known data and refers to "Walk", "Shop", and "Clean" in the above diagram. In machine learning sense, observation is our training data, and the number of hidden states is our hyper parameter for our model. Evaluation of the model will be discussed later.

State transition probabilities (A) are the arrows pointing to each hidden state. Observation probability matrix (B) are the blue and red arrows pointing to each observations from each hidden state. The matrix are row stochastic meaning the rows add up to 1.

A Second Example A bout Hidden Markov Model (HMM) is a parameterized distribution for sequences of observations. Suppose that a person (M) hears (a.k.a. observes) a sequence of T sounds o1, o2, ..., oT and he wants to reason something about this sequence. He makes the assumption that the sequence of sounds that he heard depends on a sequence of T words s1, s2, ..., sT , which he never gets to see and which is why they are called the hidden states. HMM gives a person (M) a method which, under certain assumptions, allows him to assign appropriate probabilities to sound sequences O's and word sequences S's and to make reasonable deductions about them.

HMM makes several assumptions about the data it models:

1.The observations o1, o2, ..., oT come from a known, finite sets V, called the observation space.
2.The hidden states s1, s2, ..., sT come from a known, finite sets Q, called the hidden state space.
3.The HMM assigns a probability to any given sequence s1, s2, ..., sT .

A simple program to represent the first example about HMM

```
states = ('Rainy', 'Sunny')

observations = ('walk', 'shop', 'clean')

start_probability = {'Rainy': 0.6, 'Sunny': 0.4}

transition_probability = {
   'Rainy' : {'Rainy': 0.7, 'Sunny': 0.3},
   'Sunny' : {'Rainy': 0.4, 'Sunny': 0.6},
   }

emission_probability = {
   'Rainy' : {'walk': 0.1, 'shop': 0.4, 'clean': 0.5},
   'Sunny' : {'walk': 0.6, 'shop': 0.3, 'clean': 0.1},
   }

from hmmlearn import hmm
import numpy as np

model =hmm.MultinomialHMM(n_components=2)
model.startprob_ = np.array([0.6, 0.4])
model.transmat_ = np.array([[0.7, 0.3],
                            [0.4, 0.6]])
model.emissionprob_ = np.array([[0.1, 0.4, 0.5],
                                [0.6, 0.3, 0.1]])
```

Speech Synthesis

Speech synthesis is the artificial production of human speech. A computer system used for this purpose is called a **speech computer** or **speech synthesizer**, and can be implemented in software or hardware products. We will be considering the kind of speech synthesis that takes computer-readable text and transforms it into a speech waveform. In order to perform this task accurately, the synthesis application needs to transform text into a linguistic representation that can then be converted into the acoustic domain. For example, if we want the synthesizer to pronounce the English word thought, we convert it into a **phonological representation**, such as / Θət/, and then the synthesizer produces a waveform that (hopefully) sounds like / Θət/.

A **text-to-speech** (**TTS**) system converts normal language text into speech; other systems render symbolic linguistic representations like phonetic transcriptions into speech.

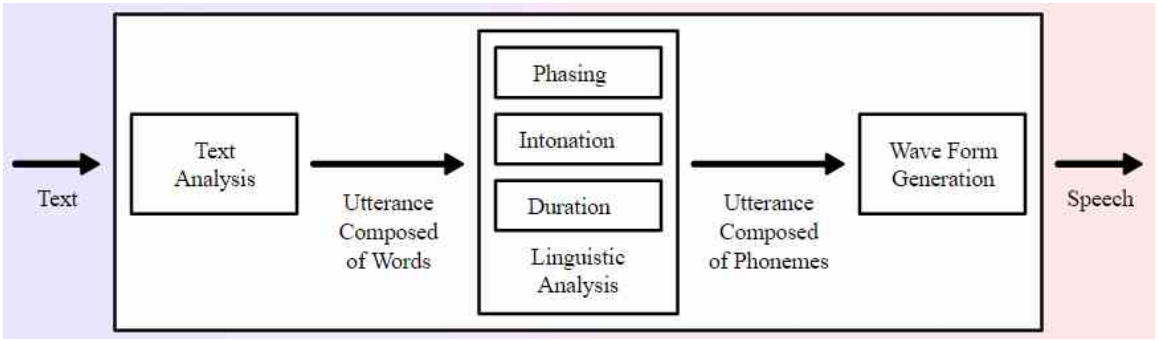


Figure 2-Overview of a typical TTS system

One common approach is **concatenative synthesis**, which takes acoustic subword units (e.g. phones, diphones, demidiphones, etc.) from a database of recorded speech and strings them together while smoothing the transitions between them.

store speech for units approximating the phoneme, rather than for words or letters. Storing examples of whole words would be prohibitive in terms of computer storage.

given the amount of words a general-purpose application is likely to encounter. In addition, such a system would be at a loss when confronted with novel words, such as names or neologisms. Storing speech examples based on letters would lead to inaccuracy, due to the irregularities of English spelling consider (through, bough, cough, enough). Synthesizers are used, together with speech recognizers:

in telephone-based conversational agents that conduct dialogues with people. Blind people and people how lost their voices can gain the benefit from the speech synthesizer.

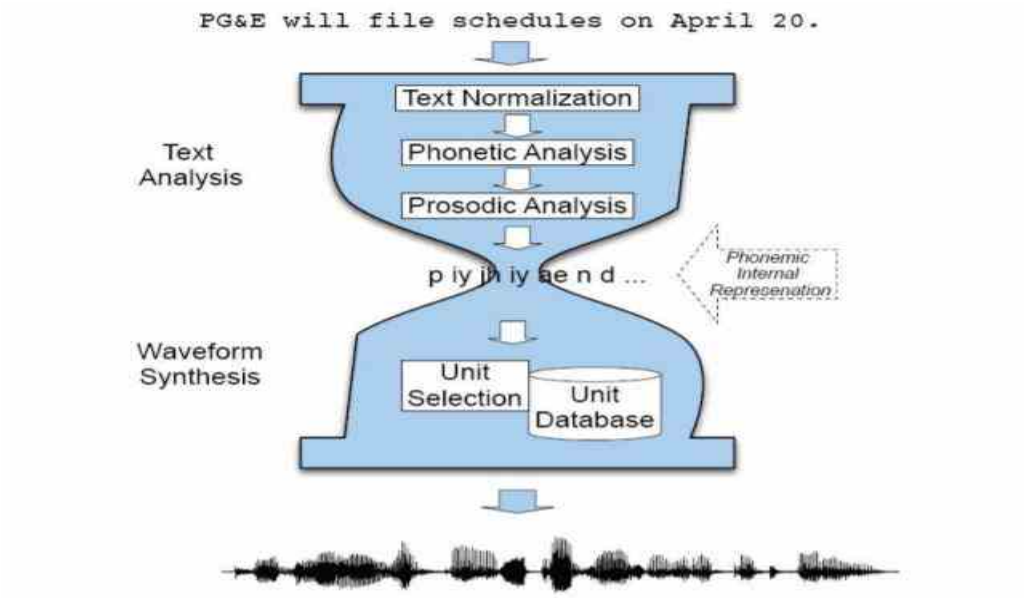
The two stages of TTS:

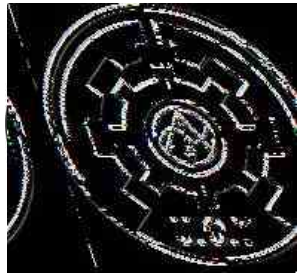
PG&E will file schedules on April 20.

1- Text Analysis: Text into intermediate representation:

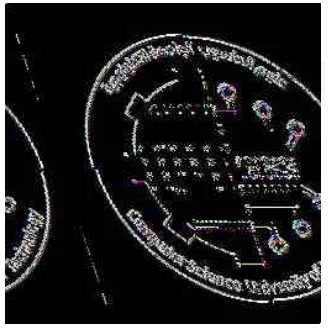
			*			*			* L-L%
P	G	AND	E	WILL	FILE	SCHEDULES	ON	APRIL	TWENTIETH
p iy	jh iy	ae n d	iy	w ih l	f ay l	s k eh jh ax l z	aa n	ey p r ih l	t w eh n t iy ax th

2- Waveform Synthesis: From the intermediate representation into waveform





Speech Recognition



2nd course lecture 6

Lecture Outlines:

- Text Normalization
- Phonetic Analysis
- Prosodic Analysis
- Machine Translation (MT)

Lecturer: Dr. Asia Ali

1. Text Normalization

The first task in text normalization is ***sentence tokenization***. In order to segment this paragraph into separate utterances for **synthesis**, we need to know that the first sentence ends at the period after March 31, not at the period of B.C. We also need to know that there is a sentence ending at the word collected, despite the punctuation being a colon rather than a period. Consider the difficulties in the following text drawn:

"He said the increase in credit limits helped B.C. Hydro achieve record net income of about $1 billion during the year ending March 31. This figure does not include any write downs that may occur if Powerex determines that any of its customer accounts are not collectible. Cousins, however, was insistent that all debts will be collected: "We continue to pursue monies owing and we expect to be paid for electricity we have sold."

The second normalization task is dealing with non-standard words. Nonstandard words include number, acronyms, abbreviations, and so on. For example, March 31 needs to be pronounced March thirty first, not March three one; $1 billion needs to be pronounced one billion dollars, with the word dollars appearing after the word billion.

. Identify tokens in text
. Chunk tokens into reasonably sized sections
. Map tokens to words
. Tag the words

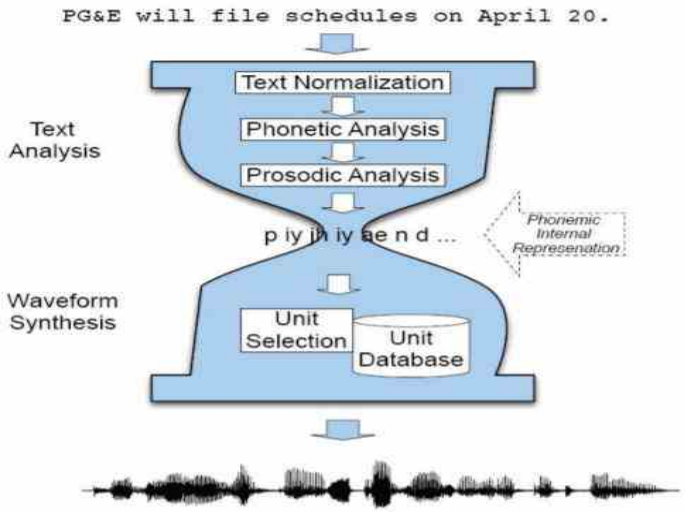


Figure -1 TTS process

2. Phonetic Analysis

The next stage in synthesis is to take the normalized word strings from text analysis and produce a pronunciation for each word. The most important component here is a large pronunciation dictionary. Dictionaries alone turn out to be insufficient, because running text always contains words that don't appear in the dictionary. Thus, the two main areas where dictionaries need to be augmented is in dealing with names and with their unknown words.

2.1 Dictionary lookup

The Dictionary lookup is the primary strategy for deriving pronunciations for input words in a text-to-speech (TTS) system. This strategy is accurate for dictionary words, but it is not complete as it is impossible to list exhaustively all input words of a language. The proper treatment of `unknown' words is currently a complex problem in TTS synthesis. Sometimes the same orthographic word has different pronunciation depending on the context of use of that word.

CMU DICTIONARY

One of the most widely- used for TTS is the freely available CMU Pronouncing Dictionary (CMU, 1993), which has pronunciations for about 120, 000 words.

The pronunciations are roughly **phonemic**, from a 39-phone ARP Abet-derived phoneme set.
Phonemic transcriptions mean that instead of marking surface reductions like the reduced vowels [ax] or [ix], CMU dict. marks each vowel with a stress tag, 0 (unstressed),1 (stressed), or 2 (secondary stress). Thus, (non-diphthong) vowels with 0 stress generally correspond to [ax] or [ix].

Most words have only a single pronunciation, but about 8,000 of the words have two or even three pronunciations, and so some kinds of phonetic reductions are marked in these pronunciations. The dictionary is not syllabified, although the nucleus simplicity marked by the (numbered) vowel.

Fig.2 shows some sample pronunciations. The CMU dictionary was designed for speech recognition rather than Synthesis uses; thus, it does not specify which of the multiple pronunciations to use for synthesis, does not mark syllable boundaries, and because it capitalizes the dictionary head words, does not distinguish between e.g., *US* and *us (the form US has the two pronunciations [AH1S] and [YUW1EH1S].*
CMU dictionary: 127K words
The CMU Pronouncing Dictionary website: http://www.speech.cs.cmu.edu/cgibin/cmudict

Phoneme	Example	Translation
AA	odd	AA D
AE	at	AE T
AH	hut	HH AH T
AO	ought	AO T
AW	cow	K AW
AY	hide	HH AY D
B	be	B IY
CH	cheese	CH IY Z
D	dee	D IY
DH	thee	DH IY
EH	Ed	EH D
ER	hurt	HH ER T
EY	ate	EY T
F	fee	F IY
G	green	G R IY N
HH	he	HH IY
IH	it	IH T
IY	eat	IY T
JH	gee	JH IY
K	key	K IY
L	lee	L IY
M	me	M IY
N	knee	N IY
NG	ping	P IH NG
OW	oat	OW T
OY	toy	T OY
P	pee	P IY
R	read	R IY D
S	sea	S IY
SH	she	SH IY
T	tea	T IY
TH	theta	TH EY T AH
UH	hood	HH UH D
UW	two	T UW
V	vee	V IY
W	we	W IY
Y	yield	Y IY L D
Z	zee	Z IY
ZH	seizure	S IY ZH ER

Figure 2- CMU sample pronunciations

Unisyn dictionary significantly more accurate, includes multiple dialects http://www.cstr.ed.ac.uk/projects/unisyn/

Dictionaries are not always sufficient: Unknown words

Big problem area is names, names are common , Spiegel (2003) estimate of US names: 2 million surnames, 100,000 first names

Personal names: McArthur, D'Angelo, Jiminez, Rajan, Raghavan, Sondhi, Zhang, Chang.

Company/Brand names: Infinit, Kmart, Cytyc, Medamicus, Inforte, Aaon, Idexx Labs, Bebe.

2.2 Letter-to-sound (grapheme-to-phoneme) (G2P)

The automatic conversion of text to phoneme is a necessary step in all-current approaches to Text-to Speech (TTS) synthesis and Automatic Speech Recognition System.

Once we have expanded nonstandard words and looked them all up in a pronunciation dictionary, we need to pronounce the remaining, unknown words.

The process of converting a sequence of letters in to a sequence of phones is called grapheme-to phoneme conversion, sometimes shortened **g2p**.

The job of a grapheme-to-phoneme algorithm is thus to convert a letter string like cake into a phone string like [KEYK].

The earliest algorithms for grapheme – to – phoneme conversion these are often called letter-to-sound or LTS rules, and they are still used in some Systems. LTS rules are applied in order, with later (default) rules only applying if the context for earlier rules is not applicable.

A simple pair of rules for pronouncing the letter c might be as follows:

c ! [k]/ {a, o} V; context-dependent

c ! [s]; context-independent

3 Prosodic analysis

Prosody basis

Prosody is the way we alternate melody and rhythm to add different meanings and emotional tone to our verbal utterances. It is the complementary to syntax and semantics – it is what gives meaning beyond structural and lexical meaning. Prosody also gives us other information about the speaker, such as gender and age.

Prosodic Analysis is the last step in the text analysis system (TTS) is prosodic analysis which provides the speech synthesizer with the complete set of synthesis controls, namely:

the sequence of speech sounds their durations

An associated pitch contour (variation of fundamental frequency with time).

The determination of the sequence of speech sounds is mainly performed by the phonetic analysis step. The assignment of duration and pitch contours is done by a set of pitch and duration rules, along with a set of rules for assigning stress and determining where appropriate pauses should be inserted so that the local and global speaking rates appear to be natural.

3.1 Prosodic structure

Often prosodic structure is described in terms of prosodic phrasing, meaning that an utterance has a prosodic phrase structure in a similar way to it having a syntactic phrase structure. For example, in the sentence: *I wanted to go to London, but could only get tickets for France.* There seems to be two main intonation phrases, their boundary occurring at the comma.

1. Prosodic phrase boundaries (often called intermediate phrases that split up the words as follows I wanted| to go| to London.

1. Practical phrase boundary prediction is generally treated as a binary classification task, where we are given a word and we have to decide whether to put a prosodic boundary after it. A text-only alternative is often used. In this method, a human labeler looks only at the text of the training corpus, ignoring the speech. The labeler marks any juncture between words where they feel a prosodic boundary might legitimately occur if the utterance were spoken.

3.2 Prosodic prominence

- In any spoken utterance, some words sound more prominent than others do. Prominent words are perceptually more salient to the listener; speakers make a word more salient in English by saying it louder, saying it slower (so it has a longer duration). We generally capture the core notion of prominence by associating a linguistic marker with prominent words, a marker called pitch accent. Words which are prominent are said to bear (be associated with) a pitch accent. Pitch accent is thus part of the phonological description of a word in context in a spoken utterance. Pitch accent is related to stress. The stressed syllable of a word is where pitch accent is realized.
- In other words, if a speaker decides to high light a word by giving it a pitch accent, the accent will appear on the stressed syllable of the word.

3.3 Tune

- Two utterances with the same prominence and phrasing patterns can still differ prosodic by having different tunes. The tune of an utterance is the rise and fall of its F0 overtime. Avery obvious example of tune is the difference between statements and yes-no questions in English. The same sentence can be said with a final rise in F0 to indicate a yes- no- question, or a final fall in F0 to indicate a declarative intonation.
- Fig.4 shows the F0 track of the same words spoken as a question or a statement. Note that the question rises at the end; this is often called a question rise. The falling intonation of the statement is called a final fall

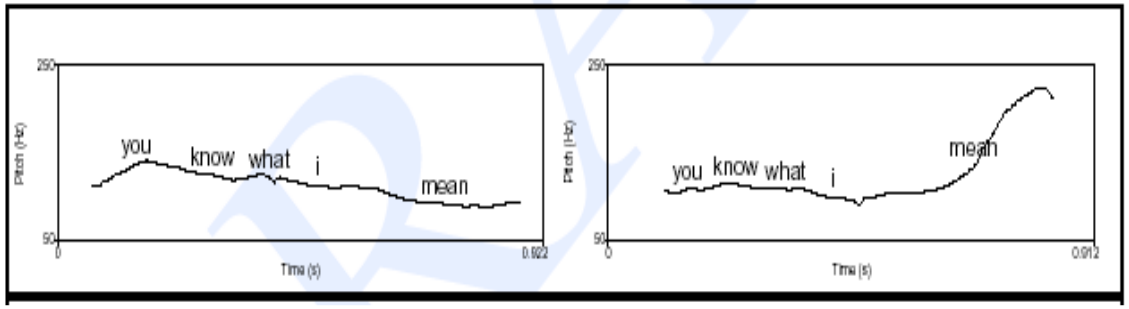


Figure 4. Tune- **The same text read as the statement you know what I mean. (on the left) and as a question You know what I mean? (on the right).**

Machine Translation (MT)

The term 'machine translation' (MT) refers to computerized systems responsible for the production of translations with or without human assistance.

Google translate instantly translates between any pair of over eighty human languages like French and English.
How does it do that?
Why does it make the errors that it does?
How can you build something better?
Modern translation systems like Google Translate *learn* to translate by reading millions of words of already translated text.

Machine Translation types

1. Direct translation
2. Transfer approaches
3. Interlingua approaches.

Machine Translation discussed in three categories:

1. Statistical Machine Translation (SMT),
2. Machine Translation using Semantic.
3. Neural Machine Translation (NMT)