"The three bhuvanas such as svrga, martya and pAtala would have been unlit and the universe would not have been lightened if and only if there was no light as of ‘indivisible shabda’, the language. This is an illustrious verse of a distinguished Indian poet Dandi written in kAvyadarsha."

Speech is the primary means of communication between people. Speech synthesis, the automatic generation of speech waveform, has been under development for several decades (Santen et al. 1997, Kleijin et al 1998). Recent progress in speech synthesis has produced synthesizers with very high intelligibility but the sound quality and the naturalness still remains a major problem. However the quality of the present synthesizers has reached an adequate level for several applications, such as multimedia and telecommunication technology. With the help of some audiovisual information or facial animation it is possible to increase speech intelligibility considerably. Speech generation is the process, which allows the transformation of a string of phonetic and prosodic symbols into a synthetic speech signal. The quality of the result is a function of the quality of the string, as well as of the quality of the generation process itself.

At the outset Text-To-Speech synthesis system consists of two main phases. The first one is text analysis, where the input text is transcribed into a phonetic and some other linguistic representation. The second one is the generation of speech waveform, where the acoustic output is produced from this phonetic and prosodic information. These two levels are normally called as the high and low level synthesis as shown in the Figure 1.1.
The input text might be for example data from a word processor, standard ASCII from e-mail, a mobile text message or a text output of a standard scanned document from a newspaper etc. The character string is then processed and analysed into phonetic representation, which is usually a string of phonemes with some additional information for correct intonation, duration and stress. Speech sound is finally generated with a synthesizer by the information from the linguistic and paralinguistic knowledge input. For an efficient TTS system usually two quality criteria are proposed. The first one is intelligibility, which can be measured by taking into account several kinds of units (phonemes, syllables, words, phrases). The second one, more difficult to define, is often labelled as pleasantness or naturalness. Actually the concept of naturalness may be related to the concept of realism in the field of image synthesis: the goal is not to destitute the reality but to suggest it. Thus, listening to a synthetic voice must allow the listener to attribute this voice to some pseudo-speaker and to perceive some kind of expressivities as well as some indices characterizing the speaking style and the particular situation of elocution. For this purpose the corresponding extra-linguistic information must be supplied to the system.

The simplest way to produce synthetic speech is to play long pre-recorded samples of natural speech, such as single word or sentences. This concatenation method provides high quality speech with naturalness, but has a limited vocabulary and usually only one voice. This method is very useful for many domain specific applications like announcement systems and information systems. However, it is quit clear that we cannot create a database of all words and common names in the world. It may be inappropriate to call it as a speech synthesizer because it contains only pre-recorded words or sentences. Thus for unrestricted speech synthesis (Text-To-Speech)
some intelligible units of speech signal is to be chosen, such as syllables, phonemes or even shorter segments.

Another widely accepted model for designing a speech synthesizer is the formant synthesizer. It is based upon the source filter model of speech production. The Figure 1.2 describes the model in detail. This method is sometimes called the terminal analogy because it models only the sound source and the formant frequencies. No physical characteristic of the vocal tract is modelled in this method of generation [89][33]. The source excitation signal could be either voiced with fundamental frequency $f_0$ or the unvoiced noise. A mixed excitation of these two may also be used for voiced consonants and some aspiration sounds. The excitation is then gained and filtered with a vocal tract filter, which is constructed of resonators similar to the formants of natural speech.

"Our objective should be to generate artificial speech by modelling the humane speech production system". [5] This is somewhat called the articulatory synthesis and it involves models of the human articulators and vocal cords. The articulators are usually modelled with a set of area function of small tube sections. The vocal cord model is used to generate an appropriate excitation signal, which may be for example a two mass model with two vertically moving masses. Articulatory synthesis holds a promise of high quality synthesised speech, but due to its complexity the potential has not been realised properly. [34][75]

All synthesis methods have their own pros and cons and sometimes it is very difficult to choose the best out of the lot. But with concatenative and formant synthesis some promising results have come to light. Articulatory method of synthetic speech generation will
definitely take a potential field in future. Different methods and algorithms are discussed in following chapters in details.

Most of the present TTS systems produce an acceptable level of intelligibility, but the naturalness dimension, the ability to control expressivities, speech style and pseudo-speaker identity still are poorly mastered. It can be mentioned however that users demands vary to a large extent according to the field of application: general public applications such as telephonic information retrieval need maximal realism and naturalness, whereas some applications involving professionals (process or vehicle control) or highly motivated persons (visually impaired, applications in hostile environments) demand intelligibility with the highest priority.

In the past few decades, various researchers have worked in the area of Speech Synthesis and Recognition and have developed different algorithms and methodologies for different speech technology development. In area of Speech Synthesis there are a number of different methodologies like Formant, Articulatory, Sinusoidal and Concatenative Synthesis.

**Formant Synthesis:** Formant synthesis is generation of speech by an acoustic-phonetic production model, based on formant of the vocal tract. A Formant Synthesizer models the speech production system in terms of the acoustic-phonetic parameters such as energy, pitch and resonance (formant) frequencies associated with speech and uses heuristic rules to derive the model.

**Articulatory Synthesis:** Articulatory synthesis tries to model the human vocal organs as perfectly as possible, so it is potentially the most satisfying method to produce high quality synthetic speech. Articulatory synthesizer generates speech by mathematically modelling
the movement of articulators, e.g. lips, tongue, and jaws. Mathematical optimisation problems are intricate.

**Sinusoidal Model based Synthesis:** Sinusoidal model are based on a well known assumption that the speech signal can be represented as a sum of sine waves with time varying amplitude and frequencies.

**Concatenative Synthesis:** In concatenative synthesis, speech is generated by combining splices of pre-recorded natural speech. To take care of context-dependency and information embedded in transition segments, the splices are selected such that they begin and end with comparatively steady states.

In the last few years there has been a significant trend for development of speech synthesizers using Concatenative based Synthesis techniques. There are a number of different methodologies for Concatenative Synthesis like TDPSOLA, PSOLA, MBROLA and Epoch Synchronous Non Over Lapping Add (ESNOLA) etc.

ESNOLA technique provides the complete control on implementation of intonation and prosody. It allows judicial selection of signal segment so that smaller fundamental parts of the phonemes may be used as units reducing both the number and the size of the signal elements in the dictionary. Further the methodology of concatenation provides adequate processing for proper matching between different segments during concatenation.

**1.1. History and Development of Speech Synthesis**

Generation of artificial speech is a dream of the mankind from a century. In this chapter a sketch of the history of synthesised speech from the first mechanical effort to systems that form the basis for today's high quality synthesiser has been drawn. Some separate
milestones in synthesis related methods and techniques are also discussed.

1.1.1. Mechanical-To-Electrical Synthesis

The methods of generation of speech through mechanical means started around two hundred years back. In St. Petersburg 1779 Russian Professor Christian Kratzenstein explained physiological differences between five long vowels (/a/, /e/, /i/, /o/, /u/) and made apparatus to produce them artificially. He constructed acoustic resonator similar to the human vocal tract and activated the resonators with vibrating reeds like in music instrument. The basic structure is shown in the Figure 1.3. The sound /i/ is produced by blowing into the lower pipe without a reed causing the flute like sound.

In 1791 Vienna and Wolfgang von Kempelen introduced the "Acoustic Mechanical Speech Machine", which was able to produce single sounds and some sound combinations (Klatt 1987, Schroeder 1993). Actually Kempelen started his work before Krazenstein (1769) and after over 20 years of research he published a book in which he described his studies on human speech production and the experiments with his speaking machine. The essential parts of the machine were a pressure chamber simulating lungs, a vibrating reed to act as vocal cords and a leather tube for simulating the vocal tract action. By manipulating the shape of the leather tube he could produce different vowel sounds. Consonants are simulated by four separate constricted passages and controlled by the fingers. For plosive sound he also employed a model of a vocal tract that include the hinged tongue and movable lips. His studies laid to the theory that the vocal tract (a cavity between the vocal cord and lips) is the main site of acoustic articulation. Before Kempelen's demonstrations the larynx was generally considered as a centre of speech production. Unfortunately,
the main mechanism of the machine was concealed, legless chess player expert. Therefore his real speaking machine was not taken seriously as it should have.

1.1.2. Development of Electrical Synthesizer

The first electrical synthesizer was introduced by Stewart in 1922 (Klatt 1987). The synthesizer had a buzzer as excitation and two resonant circuits to model the acoustic resonances of the vocal tract. The machine was able to generate single static vowel sounds with two lowest formants, but not any consonants or connected utterances. The device consisted of four electrical resonators connected in parallel and it was excited by a buzz-like source. The outputs of the four resonators were combined in the proper amplitudes to produce vowel spectra. In 1932 Japanese researchers Obata and Teshima discovered the third formant in vowels [24][51]. The first three formants are generally considered to be enough for intelligible synthetic speech. First device to be considered as a speech synthesizer, was VODER (Voice Operating Demonstrator) introduced by Homer Dudley. VODER was inspired by VOCODER (Voice Coder) developed by Bell Laboratory in mid thirties. The original VOCODER was a device for analysing speech into slowly varying acoustic parameters that could then drive a synthesizer to reconstruct the approximation of the original speech signal. The VODER consisted of wrist bar for selecting a voicing or noise source and a foot pedal to control the fundamental frequency. The source signal was routed through ten bandpass filters whose output levels were controlled by fingers. It took considerable skill to play a sentence on the device. The speech quality and intelligibility were far from good but the potential for producing artificial speech were well demonstrated. After the demonstration of VODER the scientific
world became more and more interested in speech synthesis. It was finally shown that the intelligible speech could be produced artificially. The basic structure and idea of VODER is very similar to present systems, which are based on the source-filter model of speech.

About a decade later, in 1951, Franklin Cooper and his associates developed a Pattern playback synthesizer at the Haskins laboratories. It reconverted recorded spectrogram patterns into sounds, either in original or modified form. The spectrogram patterns were recorded optically on the transparent belt.

The first formant synthesizer, PAT (Parametric Artificial Talker) was introduced by Walter Lawrence in 1953 (Klatt 1987). PAT consisted of three electronic formant resonators connected in parallel. The input signal was either a buzz or noise. A moving glass slide was used to convert painted patterns into six time functions to control the three-formant frequencies, voicing amplitude, fundamental frequency and noise amplitude. At about the same time Gunnar Fant introduced the first cascade formant synthesizer OVE-I (Orator Verbis Electris) which consisted of formant resonators connected in cascade. In 1962, Fant and Martony introduced an improved OVE-II synthesizer, which consisted of separate parts to model the transfer function of the vocal tract for vowels, nasals and obstruent consonants. Possible excitations were voicing, aspiration noise and friction noise. The OVE projects were followed by OVE-III and GLOVE at the Kungliga Tekniska Hogskolan (KTH), Sweden and the present commercial Infovox system is originally descended from these (Carlson et. al. 1981, Barber et. al. 1989, Karlsson et. al. 1993).

George Rosen at the MIT introduced first articulatory synthesizer during 1958. The DAVO (Dynamic Analog of the Vocal tract) was controlled by tap recording of control signal created by hand.
In mid 60s, first experiments with Linear Predictive Coding (LPC) were made. Linear prediction was first used in low-cost systems such as T1 Speak 'n' Spell in 1980 and its quality was quite poor compared to the present systems. But from many angles the method has been found very useful and it is used in many present systems.

The first Text-To-Speech system for English was developed in the Electro-technical Laboratory, Japan in 1968 by Noriko Umeda and his companions (Klatt 1987). It was based on an articulatory model and included a syntactic analysis module with sophisticated heuristics. The speech was quite intelligible but monotonous and far away from the quality of present systems.

In 1979 Allen, Hunnicutt and Klatt demonstrated the MITalk laboratory Text-To-Speech system developed at MIT. Two years later Dennis Klatt introduced his famous Klattalk system, which used a new sophisticated voicing source.

1.2. Theory of Speech Representation and Production

Speech processing and language technology contains lots of special concepts and terminology. To understand how different speech synthesis and analysis methods work we must have some knowledge of speech production, articulatory phonetics and some other related terminology. This following topics discusses the theory behind these terminologies.

1.2.1. Representations and Analysis of Speech Signal

Continuous speech is a set of complicated audio signals, making it difficult producing artificially. Speech signals are usually considered as voiced or unvoiced, but in some cases they are between these two. Voiced sound consists of fundamental frequency (f₀) and harmonic components produced by vocal cords (vocal folds). The
vocal tract modifies the excitation signal causing formants (pole) and sometimes antiformants (zeros) frequencies (Witten 1982). Each formant frequency has also spectral density and bandwidth and it may be sometimes difficult to define some of these parameters correctly. The fundamental frequencies and the formant frequencies are probably the most important concepts in speech synthesis and also in speech processing in general.

Fundamental frequency is absent in excitation signal in case of purely unvoiced sounds and therefore no harmonic structure is present and the resulting excitation can be considered as white noise. The airflow is forced through a vocal tract constriction, which can occur in several places between glottis and mouth. Some sounds are produced with complete stoppage of airflow followed by a sudden release, producing an impulsive turbulent excitation often followed by a more protracted turbulent excitation. The occlusion period of voiced plosive have voicing whereas that of unvoiced ones do not have voicing (Figure 1.4). Whispering is the special case of speech. When a normally voiced phoneme is whispered there is no fundamental frequency in the excitation and the first formant frequencies produced by vocal tract are perceived.

Speech signal of the three vowels /a/, /I/ and /u/ are presented in the time and frequency domain and shown in the Figure 1.5. The fundamental frequency is around 100 Hz in all the cases and the formant frequencies F1, F2 and F3 with vowel /a/ is approximately 600Hz, 1000Hz, and 2500Hz respectively. With vowel /I/ the first three formants are 200Hz, 2300Hz and 3000Hz and with /u/ 300Hz, 600Hz and 2300Hz. The harmonic structure of the excitation is also easy to perceive from frequency domain representation. It can be seen that the first three formants are inside the normal telephone channel (300Hz to
3400Hz). So the needed bandwidth for intelligible speech is not very wide. For better quality up to 10KHz bandwidth may be used which leads to 20KHz sampling frequency. Unless the fundamental frequency is outside the telephone channel, the human hearing system is capable to reconstruct it from its harmonic components.

Another commonly used method to describe the speech signal is the spectrogram representation, which is a time-frequency-amplitude representation of the signal. The spectrogram and the time domain waveform of the Oriya word sakAla (ମୋଙ) (Morning) are presented in the Figure 1.6. Higher amplitudes are presented with darker gray-levels, so the formant frequencies and the trajectories are easy to perceive. Also spectral difference of vowels and consonants are easy to comprehend. Therefore, spectrogram is perhaps the most useful representation for speech research. From the Figure 1.6 it is easy to see that the vowels have more energy and it is focussed at lower frequencies. Unvoiced consonants have considerably less energy and it is usually focussed at higher frequencies. With voiced consonants the situation is something between these two. In the Figure 1.6 the frequency axis is in kilohertz, but it is also quite common to use an auditory spectrogram where the frequency axis is replaced with Bark or Mel scale, which is normalized for hearing properties. For determining the fundamental frequency or the pitch of speech signal cepstral analysis may be used[16]. Cepstrum is obtained by first windowing and making Discrete Fourier Transform (DFT) for the signal and then logarithmizing power spectrum and finally transforming it back to the time domain by Inverse Discrete Fourier Transform (IDFT)[87]. The procedure is shown in the Figure 1.7. Cepstral analysis provides a method for separating the vocal tract information from excitation[62][7][97].
Thus the reverse transformation can be carried out to provide smoother power spectrum known as homomorphic filtering.

Fundamental frequency or the intonation contour over the sentence is important for correct prosody and natural sounding speech. The different contours are usually analysed from vocal tract in specific situation and with specific speaker characteristics and then applied to rules to generate the synthetic speech. The fundamental frequency contour can be viewed as the composite set of hierarchical patterns shown in the Figure 1.8. The overall contour is generated by the superposition of these patterns\(^{[97,54]}\). Methods for controlling the fundamental frequency are described in following chapters.

1.2.2. Human Speech Production

The vocal organs as shown in the Figure 1.8 produce human speech. The main energy source is the lungs with the diaphragm. When a person speaks the airflow is forced through the glottis between the vocal cords and the larynx to the three main cavities of the vocal tract, the pharynx and the oral and nasal cavities\(^{[87]}\). From the oral and nasal cavity the air flow exits through the nose and mouth respectively. The V-shaped opening between the vocal cords, called the glottis is the most important source of sound in the vocal system. The vocal cord may act in several different ways during speech. The most important function is to modulate the air flow by rapidly opening and closing, causing buzzing sound from which vowels and voiced consonants are produced. The fundamental frequency of vibration depends on the mass and tension and is about 110Hz, 200Hz and 300Hz with men, women and children respectively. With stop consonants the vocal cords may act suddenly from a completely closed position in which they cut the airflow completely, to totally open position producing a light cough or a glottal stop. On the other hand, with unvoiced consonants,
such as /s/ (ɔ), they may be completely open. An intermediate position may also occur with for example phonemes like /h/. The pharynx connects the larynx to the oral cavity. It has almost fixed dimensions, but its length may be changed slightly by raising or lowering the larynx at one end and the soft palate at the other end. The soft palate also isolates or connects the route from the nasal cavity to the pharynx. At the bottom of the pharynx are the epiglottis and false vocal cords to prevent food reaching the larynx and to isolate the oesophagus acoustically from the vocal tract. The epiglottis, the false vocal cords and the vocal cords are closed during swallowing and open during normal breathing. [19][20]

From technical point of view, the vocal system may be considered as a single acoustic tube between the glottis and the mouth. Glottal excited vocal tract may be then approximated as a straight pipe closed at the vocal cords where the acoustical impedance $Z_g = \infty$ (infinity) and open at the mouth ($Z_{in}=0$). In this case the volume velocity transfer function of vocal tract is

$$V(w) = Z_{in} / Z_g = U_{in} / U_g = 1 / \cos(wl/c) \quad (1.1)$$

Here 'l' the length of the tube, 'w' is the radian frequency and 'c' is the sound velocity. The denominator is zero at frequencies $F_i = w_i/2\pi$ (pi)

$$w_i/c = (2i-1)\pi/2 \quad \text{and} \quad F_i = (2i-1)/4l \quad (1.2)$$

If $l = 17 \text{ cm}$, $V(w)$ is infinite at frequency $F_i = 500\text{Hz}$, $1500\text{Hz}$, $2500\text{Hz}$,... , which resonances at every 1 KHz starting at 500 Hz. [20][21] If the length $l$ is shorter than 17 cm, the frequency $F_i$ will be scaled by a factor $17/l$, so the vocal tract may be approximated with two or three sections of tube where the areas of the adjacent sections are quite
different and resonances can be associated within individual cavities. Vowels can be approximated with a two-tube model as shown in the Figure 1.9. If we take the example of vowel /a/, the narrower tube represents the pharynx opening into a wider tube representing the oral cavity. If assumed that both tubes have an equal length of 8.5 cm, formants occur at twice the frequencies noted earlier for a single tube. Due to acoustic coupling, formants do not approach each other by less than 200Hz. So formants F1 and F2 for /a/ are not both at 100Hz, but rather 900Hz and 1100Hz, respectively. Consonants can be approximated similarly with a three-tube model as shown in the Figure 1.10. Here the narrower middle tube models the vocal tract constriction. [35][54][58] The back and middle tubes are half wavelength resonators and the front tube is a quarter-wavelength resonator with resonances

\[ \frac{c}{2l_b}, \frac{c}{2l_c}, \frac{c}{4l_f}, \] for \( i = 1, 2, 3, \ldots \)

where \( l_b, l_c, l_f \) are the length of the back, centre and front tube respectively. With typical constriction length of 3cm the resonators occur at multiples of 5333Hz and can be ignored in applications that use less than 5KHz bandwidth. [75] The excitation signal may be modelled with a two-mass model of the vocal cords, which consists of two masses coupled with a spring and connected to the larynx by strings and dampers. [80][87]

1.2.3. Speech Synthesis

Synthesized speech can be produced by several different methods. The methods are usually classified into three groups:

- Articulatory synthesis, which attempts to model the human speech production system directly.
• Formant synthesis, which models the pole frequencies of speech signal or transfer function of vocal tract based on source-filter-model.

• Concatenative synthesis, which uses different length pre-recorded samples derived from natural speech.

1.2.3.1. Articulatory Synthesis

Articulatory synthesis tries to model the human vocal organs as perfectly as possible, so it is potentially the most satisfying method to produce high-quality synthetic speech. On the other hand, it is also one of the most difficult methods to implement and the computational load is also considerably higher compared with other common methods. Thus, it has received less attention than other synthesis methods and has not yet achieved the same level of success.

Articulatory synthesis typically involves models of the human articulators and vocal cords. The articulators are usually modelled with a set of area functions between glottis and mouth. The first articulatory model was based on a table of vocal tract area functions from larynx to lips for each phonetic segment. [58] For rule based synthesis the articulatory control parameters may be for example lip aperture, lip protrusion, tongue tip height, tongue tip position, tongue height, tongue position and velic aperture are taken into consideration. Phonatory or excitation parameters may be glottal aperture, cord tension and lung pressure.

While speaking, the vocal tract muscles cause articulators to move and change shape of the vocal tract, which causes different sounds. The data for articulatory model is usually derived from X-ray analysis of natural speech. However, this data is usually only 2-D when the real vocal tract is naturally 3-D, so the rule based articulatory synthesis is very difficult to optimise due to the unavailability of
sufficient data of the motions of the articulators during speech. Other
deficiency with articulatory synthesis is that X-ray data do not
characterise the masses or degrees of freedom of the articulators. Also
the movement of tongue are so complicated that it is also impossible to
model them precisely.

1.2.3.2 Formant Synthesis

Since last decade probably it is the most widely used method
for the generation of synthetic sound. It is based upon the source-filter
model of speech. Generally there are two basic structures, parallel and
cascade. But for some better result the combination of these two are
used. Formant synthesis provides infinite number of sounds, which
makes it more flexible than the concatenation methods.

At least three formants are generally required to produce
intelligible speech and up to five formants to produce high quality
speech. Each formant is modelled with a two-pole resonator, which
enables both the formant frequencies (pole-pair frequency) and its
bandwidth is to be specified. [30][58]

Rule based formant synthesis is based on a set of rules used
to determine the parameters necessary to synthesise a desired
utterance using a formant synthesizer. The input parameters are as
follows

- Voicing fundamental frequency ($F_0$)
- Voiced excitation open quotient ($O_q$)
- Degree of voicing in excitation ($V_0$)
- Formant frequencies and amplitudes ($F_1...F_3$ and $A_1...A_3$)
- Frequency of an additional low frequency resonators ($F_N$)
- Intensity of low and high frequency region ($A_{LF}, A_{HF}$)
Here the open quotient means the ratio of the open glottis time to the total period duration.

A cascade formant synthesizer consists of band-pass resonators connected in series and the output of each formant resonator is applied to the input of the following one. The cascade structure needs only formant frequencies as control information. The main advantage of the cascade structure is that the relative formant amplitude for vowels does not need individual controls.

The cascade structure has been found better for non-nasal voiced sounds and because it needs less control information than parallel structure, it is than simpler to implement. However with the cascade model the generation of fricatives and plosive bursts is a problem. [58][93] (Figure 1.11)

A parallel formant synthesizer consists of resonators connected in parallel. Sometimes extra resonators for nasals are used. The excitation signal is applied to all formants simultaneously and their outputs are summed. Adjacent outputs of formant resonators must be summed in opposite phase to avoid unwanted zeros or anti-resonances in frequency response. The parallel structure enables controlling of bandwidth and gain for each formant individually and thus needs also more control information. The parallel structure has been found to be better for nasals, fricatives and stop consonants, but some vowels cannot be modelled with parallel formant synthesizers as well as the cascade ones. (Figure 1.12)

There has been widespread controversy over the quality and suitably characteristics of these two structures. It is easy to see that good results phowith only one basic method is difficult to achieve. So some efforts have been made to improve and combine these basic models. Klatt proposed a more complex formant synthesizer which
incorporated both the cascade and parallel synthesizers with additional resonances and anti resonances for nasalised sounds, sixth formant for high frequency noise, a bypass path to give a flat transfer function and a radiation characteristics. The system used quite complex excitation model, which was controlled by 39 parameters, updated every 5ms. The quality of Klatt formant synthesiser is very promising and the model has been incorporated into several present TTS system.[87][83]

1.2.3.3. Concatenative Synthesis

Connecting pre-recorded natural utterances is probably the easiest way to produce intelligible and natural sounding synthetic speech. However, concatenative synthesizers are usually limited to one speaker and one voice and usually require more memory capacity than other methods.

One of the most important aspects of in concatenative synthesis is to find correct unit length. The selection is usually a trade off between longer and shorter units. With longer units high naturalness, less concatenation points and good control of coarticulation are achieved, but the amount of required units and memory is increased. With shorter units, less memory is needed, but the sample collecting and labelling procedures become more difficult and complex. In present systems units used are usually words, syllables, demisyllables, phonemes, diphones and sometimes even triphones.

Word is perhaps the most natural unit for written text and some messaging systems with very limited vocabulary. Concatenation of words is relatively easy to perform and the coarticulation effects within a word are captured in stored units. However, there is a great difference with words spoken in isolation and in continuous sentence,
which makes the continuous speech to sound very unnatural. Because
there are hundreds of thousands of different words and proper names
in each language, word is not a suitable unit for any kind of unrestricted
TTS system.\textsuperscript{[30][83]}

The number of different syllables in each language considerably smaller than the number of words, but the size of unit
database is usually still too large for TTS system. For example there
are about 10,000 syllables in English. Unlike the words, the
coarticulation effect is not included in the stored units, so using
syllable, as a basic unit is not very reasonable. There is also no way to
control prosodic contours over the sentences. At the moment no word
or syllable based full TTS system exists. The current synthesis systems
are mostly based on using phonemes, diphones, demisyllables or
some kind of combination of these.

Demisyllables represent the initial and final parts of the
syllables. One advantage of demi syllables is that only about 1000 of
them is needed to construct the 10000 syllables of English.\textsuperscript{[5][30]} Using
demisyllables, instead of, for example phonemes and diphones,
requires considerably less concatenation points. Demisyllables also
take account of most transitions and then also a large number of
corticulation effects and also covers a large number of allophonic
variations due to separation of initial and final consonant clusters.
However the memory requirement is still quite high, but tolerable.
Compared to phonemes and diphones, the exact number demisyllables
in a language cannot be defined. With purely demisyllables based
systems, all possible words cannot be synthesized properly. However,
demisyllables and syllables may be successfully used in a system,
which uses variable length units and affixes.
Phonemes are probably the most commonly used units in speech synthesis because they are the normal linguistic presentation of speech. The inventory of basic units is usually between 40 and 50, which is usually the smallest compared to the other units. Using phonemes gives maximum flexibility with the rule-based system. However, some phones that do not have a steady state target position, such as plosives, are difficult to synthesize. The articulation must also be formulated as rules.\[28\][29]

Diphones (or dyads) are defined to extend the central point of the steady state part of the phone to the central point of the following one, so they contain the transition between adjacent phones. That means the concatenation point will be in the most steady state region of the signal, which reduces the distortion from concatenation point. Another advantage with the diphone is that the coarticulation effect needs no more to be formulated as rules. In principle, the number of diphones is the square of the number of phonemes (plus allophones), but not all the combinations of phonemes are needed. Though the memory requirement is bit more still the number of data is tolerable and with other advantages, diphone is a very suitable unit for sample based Text-To-Speech synthesis.\[30\][9]

Longer segmental units such as triphones or tetraphones are quite rarely used. Triphones are like diphones, but contains one phone between steady state points (half phoneme-phoneme-half phoneme). In other words, a triphone is a phoneme with a specific left and right context,

Building the unit inventory consists of three main phases.\[30\][31] First, the natural speech must be recorded so that all used units (phonemes) within all possible contexts (allophones) are included. After this, the unit must be labelled or segmented from spoken speech data,
and finally, the most appropriate units must be chosen. Gathering the samples from the natural speech is usually very time consuming. However, some of these works is done automatically by choosing the input text for analysis phase properly. The implementation of rules to select correct samples for concatenation must also be done very carefully.

The problems in concatenative synthesis compared to other methods are as follows.

- Distortion from discontinuities at concatenation points, which can be reduced using diphones or some special methods for smoothening signal.
- Memory requirements are very high, especially when long concatenation units are used, such as syllables or words.
- Data collecting and labelling of speech samples is usually time-consuming. In theory, all possible allophones should be included in the material, but trade-offs between the quality and the number of samples must be made.

Some of the problems may be solved with methods described below and the use of concatenative method is increasing due to better computer capabilities.\textsuperscript{[30][15]}

**PSOLA**

Pitch Synchronous Overlap Add (PSOLA) method was originally developed at France Telecom (CNET). It is actually not a synthesis method itself but allows pre-recorded speech samples smoothly concatenated and provides good controlling for pitch and duration, so it is used in some commercial synthesis systems, such as ProVerbe and HADIFIX. There are several versions of the PSOLA algorithm and all of them work in essence the same way. TD-PSOLA,
is the most commonly used due to its computational efficiency. The basic algorithm consists of three steps. The analysis step where the original speech signal is first divided into separate, but often overlapping Short-Term analysis signal (ST), next step the modification of each analysis signal to synthesis signal, and then the synthesis step where these segments are recombined by means of overlap-adding. Short term signals $x_m(n)$ are obtained from digital speech waveform $x(n)$ by multiplying the signal by a sequence of pitch-synchronous analysis window $h_m(n)$.

$$x_m(n) = h_m(t_m-n)x_n$$

(1.3)

Where $m$ is an index for the short-time signal. The windows, which are usually hamming type, are centred on the successive instants $t_m$, called pitch-marks. These marks are set at a pitch-synchronous rate on the voiced parts of the signal at a constant rate on the unvoiced parts. The used window length is proportional to local pitch period and the window factor is usually from 2 to 4. The pitch markers are determined either by manually inspection of speech signal or automatically by some pitch estimation methods (Korekaas et al. 1997). The segment ecombination in synthesis step is performed after defining a new pitch-mark sequence.

Manipulation of fundamental frequency is achieved by changing the time intervals between pitch markers. The modification of duration is achieved by either repeating or omitting the speech segments. In principle, modification of fundamental frequency also implies a modification of duration. Another variations of PSOLA, frequency domain PSOLA (FD-PSOLA) and Linear-Predictive PSOLA (LP-PSOLA) are theoretically more appropriate approaches for pitch scale modification because they provide independent control over the spectral envelope of the synthesis signal. FD-PSOLA is used only
for pitch scale modifications and LP-PSOLA is used with residual excited vocoders.

Some drawbacks with the PSOLA method exist. The pitch can be determined only for voiced sound and if applied to unvoiced signal parts it might generate a tonal noise.\

**Linear Prediction Based Method**

Linear predictive methods are originally designed for speech coding systems, but may be also used in speech synthesis. In fact the first synthesizers were developed from speech coders. Like formant synthesis, the LPC synthesizer is based on the source-filter model of speech. The digital filter coefficients are estimated automatically from a frame of natural speech.

The basis of linear prediction is that the current speech sample $y(n)$ can be approximated or predicted from a finite number of previous $p$ samples $y(n-1)$ to $y(n-k)$ by a linear combination with small error term $e(n)$ called residual signal.

Thus,

$$y(n) = e(n) + \sum_{k=1}^{p} a(k)y(n-k)$$  \hspace{1cm} (1.4)

$$e(n) = y(n) - \sum_{k=1}^{p} a(k)y(n-k) = y(n) - y'(n)$$  \hspace{1cm} (1.5)

where, $y'(n)$ is a predicted value, $p$ is the linear prediction order, and $a(k)$ is the linear prediction coefficients which are found by minimizing the sum of the squared errors over a frame. Two methods, the covariance method and the autocorrelation method, are commonly
used to calculate these coefficients. Only with the autocorrelation method the filter is guaranteed to be stable.\textsuperscript{[35][62]}

In synthesis phase the used excitation is approximated by a train of impulses for voiced sounds and by random noise for unvoiced. The excitation signal is then gained and filtered with a digital filter for which the coefficients are \(a(k)\). The filter order is typically between 10 and 12 at 8KHz sampling rate, but for higher quality at 22 kHz sampling rate, the order needed is between 20 and 24. The coefficients are usually updated every 5-10ms.

The main drawback in the ordinary LP method is that it represents an all pole model, which means that phonemes that contain antiformants such as nasals and nasalised vowels (e.g. /N/, /n/, /m/ etc.) are poorly modeled. The quality is also poor with short plosives because the time scale events may be shorter than the frame size used for analysis. With these deficiencies the speech synthesis quality with standard LPC method is generally considered poor, but with some modifications and extensions for the basic model the quality may be increased.

Warped Linear Prediction (WLP) takes advantage of human hearing properties and the needed order of filter is then reduced significantly from orders 20-24 to 10-14 with 22KHz sampling rate (Karjalainen et al. 1998). The basic idea is that the unit delays in digital filter are replaced by following all-pass sections

\[
\tilde{z}^{-1} = D_1(z) = \frac{z^1 - \lambda}{1 - \lambda z^{-1}}
\]  \hspace{1cm} (1.6)

Where \(\lambda\) is a warping parameter between \(-1\) and 1 and \(D_1(z)\) is a warped delay element and with Bark scale it is \(\lambda = 0.63\) with sampling rate of 22KHz. WLP provides better frequency resolution at low
frequencies and worse at high frequencies. This is almost similar to human hearing properties (Karjalainen et al. 1998).

Several other variations of linear prediction have been developed to increase the quality of the basic method. With these methods the used excitation signal is different from ordinary LP method and the source and the filter are no longer separated. These kinds of variations are for example Multi-Pulse Linear Prediction (MLPC) where the complex excitation is constructed from a set of several pulses, residual excited linear prediction (RELP) where the error signal or the residual is used as excitation signal and the speech signal can be reconstructed exactly, and code excited linear prediction (CELP) where a finite number of excitation used are stored in a finite codebook.

**Sinusoidal Model**

Sinusoidal models are based on a well-known assumption that the speech signal can be represented as a sum of sine waves with time varying amplitudes and frequencies. In the basic model, the speech signal $s(n)$ is modeled as the sum of a small number $L$ of sinusoids

$$s(n) = \sum_{l=1}^{L} A_l \cos(\omega_l n + \phi_l)$$  \hspace{1cm} (1.7)

where $A_l(n)$ and $\phi_l(n)$ represent the amplitude and phase of each sinusoidal component associated with the frequency track $w_k$. To find these parameters $A_l(n)$ and $\phi_l(n)$, the DFT of windowed signal frames is calculated, and the peaks of the spectral magnitude are selected from each frame (Figure 1.13). The basic model is also known as the McAulay / Quatieri Model. The basic model has some modifications such as ABS/OLA (Analysis By Synthesis / Overlap Add) and Hybrid / Sinusoidal Noise models (Macon 1996).
While the sinusoidal models are perhaps very suitable for representing periodic signals, such as vowels and voiced consonants, the representations of unvoiced speech become problematic.

1.2.3.4. High Level Synthesis

With high-level synthesis the input text or information is transcribed in such format that low-level voice synthesizer is capable to produce the acoustic output. A proper implementation of this is the fundamental challenge in all present systems and will probably be for years to come. The procedure consists of three main phases and they are

- Text preprocessing where numerals, special characters, abbreviations and acronyms are expanded into full words.
- Pronunciation analysis where the pronunciation of certain words including homographs and proper names are determined.
- Prosodic analysis where the prosodic features of speech are determined.

This is discussed in detail in the section 1.2.4. After High-level synthesizers, the information is delivered to drive some low-level systems. The type of used data depends on the driven system. For example, for formant synthesizers, at least fundamental frequency, formant frequencies, duration and amplitude of each sound segment is needed.

1.2.3.5. Other Methods of Speech Synthesis

Several methods have been put forward to improve the quality of the output. Different methods have been studied to generate individual phoneme. Time domain synthesis can produce high quality
and natural sounding speech segments, but in some segments combinations the synthesized speech is discontinuous at the segment boundary and if a wide-range variation of fundamental frequency is required, overall complexity will increase. On the other hand, formant synthesis yields more homogeneous speech allowing a good control of fundamental frequency; but the voice timber sounds more synthetic. This approach leads to the hybrid system, which combines both the time and frequency domain

The basic idea of a hybrid system is shown in the Figure 1.14. Several methods and techniques for determining the control parameters for a synthesizer may be used. Recently ANN (Artificial Neural Network) based method is being used to control the synthesis parameters such as duration, gain and fundamental frequency. Neural networks have been applied in speech synthesis system for a decade back. The processing elements or the neurons are interconnected in a network that can identify patterns in data as it is exposed to the data.

1.2.4. Issues in High Level Speech Synthesis

1.2.4.1. Text Preprocessing

The first part of all TTS system is to convert input data to an intermediate format for a synthesizer. In this stage, all non-alphabetical characters, numbers, abbreviations and acronyms must be converted into a full spelled-out formant. Text preprocessing is usually made with simple one-to-one look up tables, but in some cases additional information of neighboring words or characters is needed. This may lead to a large database and complicated set of rules and may cause some problems with real time systems. Input text may also contain some control characters, which must be delivered through the text parser without modifications. The conversion must neither affect abbreviations, which are a part of another.
Numbers are perhaps the most difficult to convert correctly into spelled-out format. Numbers are used in several relations, such as digits, dates, roman numerals, measures and mathematical expressions. Numbers between 1100 and 2999 are usually converted as years like 1910 as nineteen-ten etc, Expression in the form 12/12/04 or 11/11/2004 may be converted as dates if the numbers are within acceptable values. However, the expression 2/5 is more difficult because it may be either two divided by five or the second of May. In some cases, the correct conversion is possible to conclude from compounding information (measures etc.) or from the length of the number (dates, phone numbers etc.). However, there will be always some ambiguous situations. 

In some cases with measures, usually currencies, the order of some characters and value is changed. For example Rs. 3.02 is converted as three rupees and two-paisa. In these situations, the numerical expressions, which are already in spelled-out format, must be recognized to avoid the misconversion like Rs. 100 thousands to one hundred rupees thousand.

Some abbreviations and acronyms are ambiguous in different context. For common abbreviations like st., the first thing to do is to check if it is followed by a capitalised word (potential name), when it will be expanded as saint. Otherwise, if it is preceded a capitalised word, an alphanumeric (5th), or a number, it will be expanded as street.

To get rid of some of these problems an efficient parser may be used by programming or a parsing database. A parsing database provides more flexibility for corrections afterwards but may have some limitations with abbreviations, which have several different versions of correct conversions. A line in the converting database may look as follows:
where the "rules" may contain information of in which cases the current abbreviation is converted, e.g. if it is accepted in capitalised form or accepted with period or colon. Preceding and following information may contain also the accepted forms or ambient text, such as number, spaces and character characteristics (vowel/consonants etc.).

Sometimes different special modes, especially with numbers, are used to make this stage more accurate. E.g. math mode for mathematical expressions and date mode for dates and so on. Another situation where the specific rules are needed is for example the email message where the header information needs more attention.

1.2.4.2. Pronunciation

Analysis for correct pronunciation from written text has also been one of the most challenging tasks in the field of speech synthesis. One method can be to store as much names as possible into a specific pronunciation table. Due to the amount of existing names, this is quite unreasonable. So rule based system with an exception dictionary for words that fail with those letters-to-phoneme rules may be a much more reasonable approach. This approach is also suitable for normal pronunciation analysis. With morphemic analysis, a certain word can be divided in several independent parts, which are considered as the minimal meaningful subpart of words as prefix, root and affix. About 12000 morphemes are located in English. However it has been seen that the morpheme analysis may fail with the word pair.

Another relatively good approach for the pronunciation problem is a method called pronunciation by analogy where a novel word is recognized as parts of the known words and the part pronunciations are built up to produce the pronunciation of a new word. E.g. pronunciation of word grip may be constructed from grin and rip.
Such kind of problems can be solved if we can design a speech mark-up language for Indian languages.

1.2.4.3. Prosody

Prosodic or the suprasegmental features consists of pitch, duration and stress over the time. With good control over the gender, age, emotions and other features in speech can be well modelled. However, almost everything seems to have effect on prosodic features of natural speech, which makes accurate modelling very difficult. Prosodic features can be divided into several levels such as syllable, word or phrase level. For example, at word level vowels are more intense than consonants. At phrase level correct prosody is more difficult to produce than at the word level. [90]

The pitch pattern or the fundamental frequency over a sentence (intonation) in natural speech is a combination of many factors. The pitch contour depends on the meaning of the sentence. For example, in normal speech the pitch slightly decreases towards the end of the sentence and when the sentence is in the question form, the pitch pattern will rise to the end of the sentence. In the end of the sentence there may also be a continuous rise, which indicates that there is more speech to come. A rise or fall in fundamental frequency can also indicate a stressed syllable. [67] Finally the pitch contour is also affected by gender, physical and emotional state and attitude of the speaker.

The duration or the time characteristics can also be investigated as several levels from phoneme (segmental) durations to sentence level timing, speaking rate and rhythm. The segmental duration is determined by a set of rules to determine correct timing. Usually some inherent duration for phoneme is modified by rules between maximum and minimum durations. For example consonant in
non-word initial positions are shortened, emphasized words are significantly lengthened, or a stressed vowel or sonorant preceded by a voiceless plosive is lengthened. In general the phoneme duration differs due to neighbouring phonemes. At sentence level, the speech rate, rhythm and correct placing of pauses for correct phrase boundaries are important. For example, a missing phrase boundary just makes speech sound rushed which is not as bad as an extra boundary, which can be confusing. With some methods to control duration or fundamental frequencies, such as PSOLA, the manipulation of one feature affects another.

The intensity pattern is perceived as a loudness of speech over time. At syllable level vowels are usually more intense than consonants and at a phrase level syllables at the end of the utterance can become weaker in intensity. The intensity pattern in speech is highly related with fundamental frequency. The intensity of a voiced sound goes up in proportion to fundamental frequency.

The speaker's feeling and emotional state affect speech in many ways and the proper implementation of these features in synthesized speech may increase the quality considerably. With text-to-speech systems this is rather difficult because written text usually contains no information of these features. However, this kind of information may be provided to a synthesizer with some specific control characters or character strings.

This section shortly introduces how some basic emotional states affect voice characteristics. The voice parameters affected by emotions are usually categorised in three main types.

- **Voice Quality**, which contains largely constant voice characteristics over the spoken utterance, such as loudness or breathiness. For example, angry voice is breathy, loud and
has a tense articulation with abrupt changes while sad voice is very quite with a decreased articulation precision.

- **Pitch contour** and its dynamic changes carry important emotional information, both in the general form for the whole sentence and in small fluctuations at word and phonemic levels. The most important pitch features are the general level, the dynamic range, changes in overall shape, content words, stressed phonemes, emphatic stress and cluse boundaries.

- **Time characteristics** contain the general rhythm, speech rate, the lengthening and shortening of the short syllables, the length of content words and the duration and the placing of pauses.

The number of possible emotions is very large, but there are five discrete emotional states, which are commonly referred as the primary or the basic emotions and the others are altered or mixed forms of these. These are anger, happiness, sadness, fear and disgust. The secondary emotional states are for example whispering, shouting, grief and tiredness.

Anger in speech causes increased intensity with dynamic changes. The voice is very breathy and has tensed articulation with abrupt changes. The average pitch pattern is higher and there is a strong downward inflection at the end of the sentence. The pitch range and its variations are also wider than in normal speech and the average speech rate is also a little bit faster.

Happiness or joy causes slightly increased intensity and articulation for content words. The voice is breathy and light without tension. Happiness also leads to increase in pitch and pitch range. The
peak values of pitch and the speech rate are the highest of basic emotions.

Fear or anxiety makes the intensity of speech lower with no dynamic changes. Articulation is precise and the voice is irregular and energy at lower frequency is reduced. The average pitch and the pitch range are slightly higher than in neutral speech. The speech rate is slightly faster than in normal speech and contains pauses between words forming almost one third of the total speaking time.

Sadness or sorrowness in speech decreases the speech intensity and its dynamic changes. The average pitch is at the same level as in natural speech, but there are almost no dynamic changes. The articulation precision and the speech rate are also decreased. High ratio of pauses to phonation time also occurs. Grief is an extreme form of sadness where the average pitch is lowered and the pitch range is very narrow. Speech rate is very slow and pauses form almost a half of the total speaking time.

Disgust or contempt in speech also decreases the speech intensity and its dynamic range. The average pitch level and the speech rate are also lower compared to normal speech and the number of pauses is high. Articulation precision and phonation time are increased and the stressed syllables in stress content words are lengthened.  

Whispering and shouting are also common versions of expressions. Whispering is produced by speaking with high breathiness without fundamental frequency, but the emotions can still be conveyed. Shouted speech causes an increased pitch range, intensity and greater visibility in it. Tiredness causes a loss of elasticity of articulatory muscles leading to lower voice and narrow pitch range.
1.2.5. Applications of Synthetic Speech

Synthetic speech may be used in several applications. Communication aids have developed from low quality talking calculators to modern 3D applications, such as talking heads. The implementation method depends mostly on used applications. In some cases, such as announcement or warning system, unrestricted vocabulary is not necessary and the best result is usually achieved with some simple messaging system.

On the other hand, some applications, such as reading machine for the blind or electronic mail readers, require unlimited vocabulary and a TTS system is needed.

The application field of synthetic field is expanding very fast, while the quality of TTS system is also increasing steadily. Speech synthesis systems are also more affordable for common man, which makes the system more suitable for everyday use.

Application for Blind

Probably the most important and useful application field in speech synthesis is the reading and communication aids for the blind. Before synthesized speech, specific audio books were used where the content of the book was read into audiotape. It is clear that making such spoken copy of any large book takes several months and is very expensive. It is also easier to get information from computer with speech instead of using special bliss symbol keyboard, which is an interface for reading the Braille characters.

The first commercial TTS application was probably the Kurzweil reading machine for the blind introduced in 1970. It consisted of an optical scanner and text recognition software and was capable to produce quite intelligible speech from written multi-font text. The prices
of the first reading machines were far too high for average user and these machines were used mostly in libraries or related places. Today, the quality of reading machines has reached an acceptable level and prices have become affordable for single individual, so a speech synthesizer will be very helpful and common device among visually impaired people in future. Current systems are mostly software based, so with scanner and OCR system, it is easy to construct a reading machine for any computer environment with tolerable expanses. Regardless of how fast the development of reading and communication aids is, there is always some improvement to do.

The most crucial factor with reading machine is speech intelligibility, which should be maintained with speaking rates ranging from less than half to at least three times normal rate. Naturalness is also an important feature and makes the synthetic speech more acceptable. Although the naturalness is one of the most important features, it may sometimes be desirable that the listener is able to identify that speech is coming from machine, so the synthetic speech should sound natural but somehow "neutral".

When the output from a speech synthesizer is listened for the first time, it may sound intelligible and pleasant. However, during longer listening period, single click or the other weak points in the system may arise very annoying. This is called annoying effect and it is difficult to perceive with any short-term evaluation method, so for these kind of cases, the feedback from long-term users is sometimes very essential.

Currently the speech synthesisers are used to read the World Wide Web (WWW) pages from different media with normal personal computers. Information services may also be implemented through a normal telephone interface with keypad control similar to "Text TV". With modern computers it is also possible to add new features into
reading aids. It is possible to implement software to read standard check forms or find the information how the newspaper article is constructed. However, sometimes it may be impossible to find correct construction of the newspaper article if it is for example divided in several pages or has an anomalous structure.

A blind person cannot also see the length of the input text when starting to listen with a speech synthesizer, so an important feature is to give in advance some information of the text to read. For example the synthesizer may check the document and calculate the estimated duration of reading and speak it to the listener. Also the information of bold or underlined text may be given by for example with slight change of intonation or loudness.

**Application for Deaf and Vocally Handicapped**

People who are born deaf cannot learn to speak properly and people with hearing difficulties have usually speaking difficulties. Synthesized speech gives the defended and vocally handicapped an opportunity to communicate with people who do not understand the sign language. With a talking head it is possible to improve the quality of the communication situation even more because the visual information is the most important with the deaf and dumb. A speech synthesis system may also be used with communication over the telephone line. Adjustable voice characteristics are very important in order to achieve individual sounding voice. Users of talking aids may also be very frustrated by an inability to convey emotions, such as happiness, sadness, urgency or friendliness by voice.

With keyboard it is usually much slower to communicate than with normal speech. One way to speed up this is to use the predicative input system that always prompt the most frequent word for any typed word fragment and then the user can then hit a special key to accept
the prediction. Even individual pre-composed phrases, such as greetings or salute can be used.

Educational applications

A computer with speech synthesizer can teach 24 hours a day and 365 days a year. It can be programmed for special tasks like spelling and pronunciation teaching for different languages. It can also be used with interactive educational applications.

Especially with people who are impaired to read (dyslexics), speech synthesis may be very helpful because especially some children may feel themselves very embarrassing when they have to be helped by a teacher. It is also almost impossible to learn writing and reading without spoken help. With proper computer software, unsupervised training for these problems is easy and inexpensive to arrange.

A speech synthesizer connected with word processor is also helpful and to proof reading. Many users find it easier to detect grammatical and stylistic problems when listening than reading. Normal misspellings are also easier to detect.

Telecommunication and Multimedia Applications

Synthesized speech has been used for decades in all kind of telephone enquiry systems, but the quality has been far from good for common customers. Today, the quality has reached the level that normal customers are adopting it for everyday use. E-mail has become very usual in last few years. However, it is sometimes impossible to read those e-mail message when being for example abroad. There may be no proper computer available or some security problem exists. When synthetic speech e-mail message may be listened to via normal
telephone line. Synthesized speech may also be used to speak out short-text-messages in mobile phones.

**Future Applications (Audio Visual Speech Synthesis)**

Practically the speech synthesis system can be used in all kind of human machine interactions. In warning or the alarm system synthesized speech may be used to give more accurate information of the current situation. Using speech instead of warning lights or buzzers gives an opportunity to reach the warning signal from different rooms. Speech synthesizers may also be used to receive some desktop messages from a computer, such as printer activity and received e-mail.

In future if the speech recognition technique reach adequate level, synthesized speech may also be used in language interpreters or several other communication systems, such as videophones, videoconferencing or talking mobile phones. If it is possible to recognize speech, transcribe it into ASCII strings and then resynthesise it back to speech, a large amount of transmission capacity may be saved. With talking mobile phones it is possible to increase the usability considerably for example with visually impaired persons or in situations where it is difficult or even dangerous to try to reach the visual information. It is obvious that it is less dangerous to listen than to read the output from mobile phones for example while driving the car.

During last few decades the communication aids have been developed from talking calculators to modern three-dimensional audiovisual applications. The application field for speech synthesis is becoming wider all the time, which brings also more funds into research and development areas.

Speech communication relies not only on audition, but also on visual information. Facial movements such as smiling, grinning, eye
blinking, head nodding and eyebrow rising give an important additional information of the speaker's emotional state. The emotional state may be even concluded from facial expressions without any sound. Fluent speech is also emphasized and punctured by facial expressions. With visual information added to synthesize speech it is also possible to increase the intelligibility significantly, especially when the auditory is degraded by the noise, bandwidth filtering or hearing impairment. The visual information is especially helpful with front phonemes whose articulation we can see, such as labiodentals and bilabials. For example, intelligibility between /b/ and /d/ increase significantly with visual information. Synthetic face also increases the intelligibility with natural speech. However, the facial gesture and speech must be coherent. Without coherence the intelligibility of speech may be even decreased. If an audio syllable /ba(ə)/ is dubbed into a visual /ga(ə)/, it is perceived as /da(ə)/ and this effect is known as McGurk effect.

Human facial expression has been under investigation for more than one hundred years. The first computer based modelling occurs 30 years ago. In 1972 Parke introduced the first three-dimensional face model and in 1974 he developed the first version of the parametric three-dimensional model. The development in the direction of facial animation is also very fast with the addition of the faster machine available.

The addition of the facial animation to the speech synthesizers has been applied from last ten years. Most of the audiovisual speech synthesizers are based upon the parametric model developed by Parke in 1982. The model consists of normally the collection of 1000 polygons that approximated the surface of the human face including the eyes, the eyebrows, the lips and the teeth.
The polygon surface was controlled by using 50 parameters. However, the present day systems contain a number of modifications to Parke model to improve it and to make it more suitable for synthesized speech. These are usually a set of rules for generating facial control parameter trajectories from phonetic text, and a simple tongue model, which were not included in the original Parke model.

Additional visual information is very helpful for hearing impaired people. It can be used as a tool for interactive training of speech reading. It may be used in information systems in public and noisy environment, such as airports, train stations and shopping centres. If it is possible to make the talking head look like certain individual, it may be utilized in video conferencing.

The easiest approach is to use pre-stored images to represent all the possible shapes under interest and combine these with some morphing method similar to concatenative speech synthesis. This is not a flexible method as there is no control to the different facial features independently of each other. The talking head is usually implemented with some kind of parametric model. There are usually two basic methods and they are

- Two or three-dimensional parametric model which can be viewed as geometric description of the facial surface that can be deformed using a limited set of control parameters and rendered using standard computer graphics techniques. The method is similar to formant synthesis.
- Muscle based controlling, where the face surface is modelled with facial muscle activation parameters. The method is perhaps theoretically most elegant because it models face movements directly as the articulatory synthesis models the vocal system. However, there are several difficulties in
modelling these models, as all the muscles need to be simulated.

Due to difficulties with muscle-based implementation, the parametric model seems to be more feasible. Naturally, the mouth and the lips are the most important in facial models, but with eyes, eyebrows, jaw and tongue it is possible to make the audiovisual speech more natural and intelligible. In visual part the equivalence of phonemes is called as visemes.

Like in concatenative speech synthesis, diphone like units may be used to avoid the discontinuities and to include co-articulation effect in used units. In visual parts these units are called di-visemes. A di-viseme records the change in articulation produced while moving from one visemes to another. Longer units such as tri-visemes may be used which contains the immediate right and left context effect on a centre visemes.

Audiovisual speech suffers mostly of the same problem as normal speech synthesis. For example, phoneme /t/ in tea differs in lip shape to the same phoneme in two. These differences in facial movements due to context are the visual correlate of the speech effect known as coarticulation.

The computational requirements for the visual part are usually considerably higher than for the audio, so some kind of feedback from the visual part may be needed to avoid lag between audio and video. The lag may be avoided by buffering the audio images and adjusting the frame rate if necessary.
Figure 1.1. Text-To-Speech Synthesis Procedure

Figure 1.2. Source Filter Model of Speech

Figure 1.3. Kratzenstein Resonators
Figure 1.4 Spectrogram representation of unvoiced sound (/k/ Q) and voiced sound (/g/ Q).

Figure 1.5. Time and frequency domain representation of three fundamental vowels /a/ (Q), /i/ (Q) and /u/ (Q).

Figure 1.6. Spectrogram representation of the word sakAla
Figure 1.7. Cepstral Coefficient determination (a) Complex (b) Real

Figure 1.8. Human Speech Production System

Figure 1.9. Two-tube Model

Figure 1.10. Three-tube Model
Figure 1.11 Cascade Formant Synthesizer

Figure 1.12. Parallel Format Synthesizer

Figure 1.13. Sinusoidal analysis/synthesis system
Figure 1.14. Basic Idea of the Hybrid Synthesis System

Figure 1.15. Structure of the Audio-Visual Synthesizer