*M.Tech. Credit Seminar Report, Electronic Systems Group, EE Dept, IIT Bombay, submitted November 2004*

# SPEECH SYNTHESIS BASED ON SINUSOIDAL MODELING

**K. Pavan Kumar (Roll no. 04307023)**

*(Supervisor: Prof. Preeti S. Rao)*

## ABSTRACT

This report presents an introduction to speech synthesis with a brief overview of some methods and their associated problems. The usage of sinusoidal representation of speech waveform for producing synthetic speech is discussed in detail. The sinusoidal model is based on extracting the amplitudes, frequencies, and phases of the component sine waves from the short-time Fourier transform and using them for the production of synthetic speech. The report is based on some of the works carried out on speech modification to achieve desirable characteristics in speech using pitch scaling, time-scaling, and spectral warping. Obtaining expressions like anger, emotion, joy etc., in synthetic speech is one of the major areas of concern these days. This aspect of bringing expression into speech using the sinusoidal model is presented in brief.

**Contents**

1. Introduction to speech synthesis
2. Problems in text-to-speech synthesis
3. Methods for speech synthesis: Overview
4. Speech analysis/synthesis: Sinusoidal representation
5. Discussion

References

# 1. INTRODUCTION TO SPEECH SYNTHESIS

Speech is the primary means of communication between people. Speech synthesis, automatic generation of speech waveforms, has been under development for several decades [1]. Recent progress in speech synthesis has produced synthesizers with very high intelligibility but the sound quality and naturalness still remain a major problem. However, the quality of present products has reached an adequate level for several applications, such as multimedia and telecommunications. This section gives a brief introduction to the history of speech synthesis and the basic idea of text-to-speech synthesis.

## 1.1. History of Speech Synthesis

Artificial speech has been a dream of the humankind for centuries. The earliest efforts to produce synthetic speech were made over two hundred years ago [1]. In St. Petersburg 1779 Russian Professor Christian Kratzenstein explained physiological differences between five long vowels (/a/, /e/, /i/, /o/, and /u/) and made apparatus to produce them artificially. A few years later, in Vienna 1791, Wolfgang von Kempelen introduced his "Acoustic- Mechanical Speech Machine", which was able to produce single sounds and some sound combinations. In about mid 1800's, Charles Wheatstone constructed his famous version of von Kempelen's speaking machine. It was a bit more complicated and was able to produce vowels and most of the consonant sounds. The first full electrical synthesis device was introduced by Stewart in 1922. First device to be considered as a speech synthesizer was VODER (Voice Operating Demonstrator) introduced by Homer Dudley in New York World's Fair 1939. Some milestones [1] in speech synthesis development are shown in Figure 1.1.

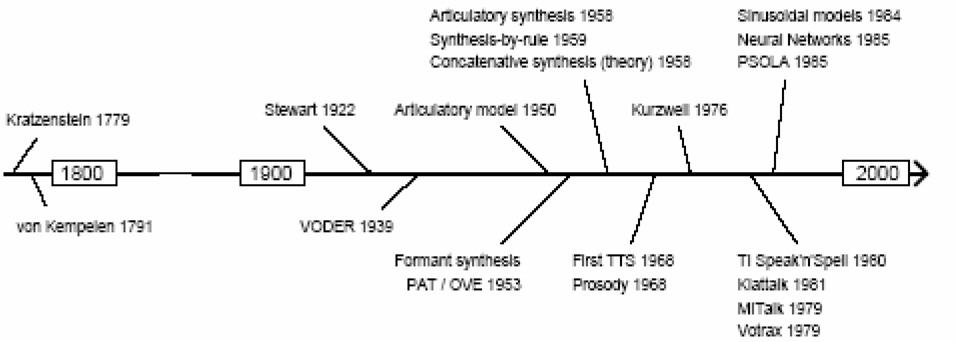

Fig.1.1. Some milestones in speech synthesis [1].

## 1.2. Text-to-Speech synthesis

The text-to-speech synthesis (TTS) procedure consists of two main phases. The first one is text analysis, where the input text is transcribed into a phonetic or some other linguistic representation, and the second one is the generation of speech waveforms, where the acoustic output is produced from this phonetic and prosodic information. These two phases are usually called as high- and low-level synthesis. In high-level synthesis the input text is converted into such form that the low-level synthesizer can produce the output speech. The three basic methods for low-level synthesis are the formant, concatenative, and articulatory synthesis [1].

A simplified version of the text-to-speech procedure is presented in Figure 1.2. The input text might be for example data from a word processor, standard ASCII from e-mail, a mobile text-message, or scanned text from a newspaper. The character string is then preprocessed and analyzed 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 the low-level synthesizer by the information from high-level one. Real-time TTS produces speech that is generally intelligible, but lacks naturalness. Quality inferior to that of human speech is usually due to inadequate modeling of three aspects of human speech production: coarticulation, intonation, and vocal tract excitation [5].

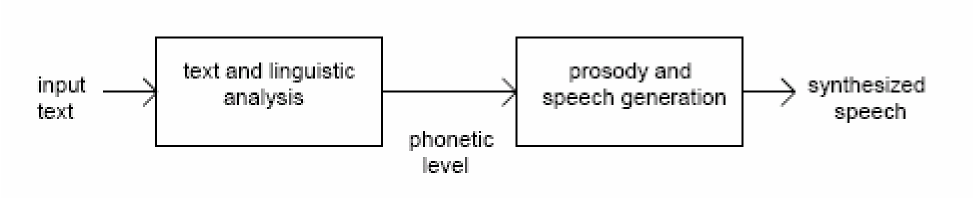

Fig.1.2 Text-to-Speech synthesis procedure [1].

The critical issues for current synthesizers concern trade-offs among the conflicting demands of maximizing speech quality, while minimizing memory space, algorithmic complexity, and computational speed [5]. While simple TTS is possible in real time with low cost hardware, there is a trend toward using more complex programs.

## 1.3. Organization of the Report

The main objective of the report is to study various methods in speech synthesis, identify their problems and choose an appropriate method for achieving expression in the synthetic speech.

Firstly in section 2, the problems involved in text-to-speech synthesis such as text preprocessing, prosody modification and speech generation are discussed. In section 3, some of the commonly used synthesis methods are presented. A study of various synthesis methods proved the sinusoidal model to be more appropriate in achieving expressive speech synthesis. Section 4 is devoted to explain the underlying principles of the sinusoidal analysis/synthesis system. This section also discusses the application of sinusoidal model to do speech transformations in achieving expression in synthetic speech.

## 2. PROBLEMS IN TEXT-TO-SPEECH SYNTHESIS

The problem area in speech synthesis is very wide. There are several problems in text pre-processing, such as numerals, abbreviations, and acronyms. Correct prosody and pronunciation analysis from written text is also a major problem. Written text contains no explicit emotions and pronunciation of proper and foreign names is sometimes very anomalous. This section presents the various problems in speech synthesis.

### 2.1 Text-to-Phonetic Conversion

The first task faced by any TTS system is the conversion of input text into linguistic representation, usually called text-to-phonetic or grapheme-to-phoneme conversion. Conversion can be divided into three main phases, text preprocessing, creation of linguistic data for correct pronunciation, and the analysis of prosodic features for correct intonation, stress, and duration.

#### *2.1.1 Text preprocessing*

Text preprocessing is usually a very complex task and includes several language dependent problems [1]. Digits and numerals must be expanded into full words. For example in English, numeral 243 would be expanded as *two hundred and forty-three* and 1750 as *seventeen-fifty* (if year) or *one-thousand seven-hundred and fifty* (if measure). Fractions and dates are also problematic. The second task is to find correct *pronunciation* for different contexts in the text. Some words, called *homographs*, cause may be the most difficult problems in TTS systems. Homographs are spelt the same way but they differ in meaning and usually in pronunciation (e.g. fair, lives). Finding correct intonation, stress, and duration from written text is probably the most challenging problem [1]. These features together are called *prosodic* features and may be considered as the melody, rhythm, and emphasis of the speech at the perceptual level. The prosodic dependencies are shown in Figure 2.1. Written text usually contains very little information of these features and some of them change dynamically during speech. However, with some specific control characters this information may be given to a speech synthesizer.

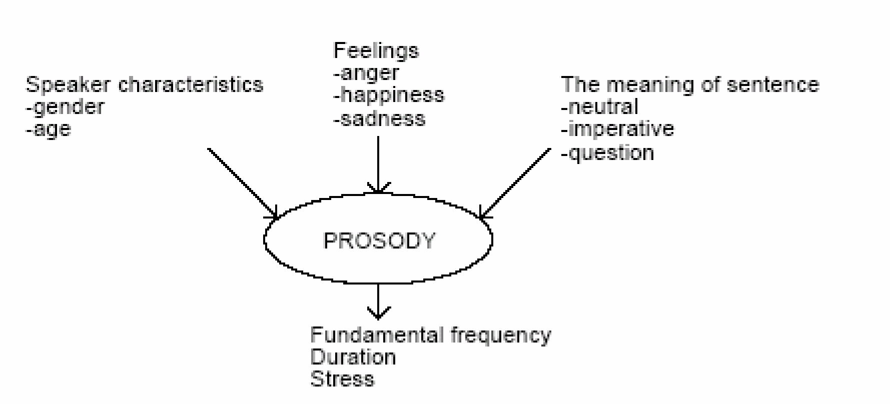

Fig. 2.1. Prosodic dependencies [1].

### 2.2. Problems in low-level synthesis

At the low-level synthesis, the discontinuities and contextual effects in wave concatenation methods are the most problematic. With some larger units, such as words or syllables, the coarticulation effect is a problem and some problems with memory and system requirements may arise [1]. The other methods to produce sounds are formant synthesis and articulatory synthesis. The formant synthesis is not consistent with nasalized sounds. With articulatory synthesis, the collection of data and implementation of rules that drive that data correctly is complex.

## 3. METHODS FOR SPEECH SYNTHESIS: OVERVIEW

Synthesized speech can be produced by several different methods. All of these have some benefits and deficiencies. 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 prerecorded samples derived from natural speech.

### 3.1 Articulatory Synthesis

Articulatory synthesis is based on modeling the human speech production system hence this approach is capable of producing high quality speech. This method is computationally more difficult when compared to other methods [1]. Articulatory models are useful to study the physics of speech production [5].

Articulatory synthesis involves models of the human articulators and vocal cords. The articulators are usually modeled with a set of area functions between glottis and mouth. 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. Phonatory or excitation parameters may be glottal aperture, cord tension, and lung pressure [1].The rule base consisting of the area functions for various sounds can be obtained by X-ray analysis of natural speech.

*Advantage*

One advantage of articulatory synthesis is that the vocal tract models allow accurate modeling of transients due to abrupt area changes [5]. Also, this is the most accurate model since it directly models the human speech production system.

*Some deficiencies*

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 optimize due to the unavailability of sufficient data of the motions of the articulators during speech.

### 3.2 Formant synthesis

Another widely used method to produce synthetic speech is formant synthesis which is based on the source-filter-model of speech production described in Figure 3.1 below. The method is sometimes called terminal analogy because it models only the sound source and the formant frequencies, not any physical characteristics of the vocal tract [1]. The excitation signal could be either voiced with fundamental frequency (F0) or 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.

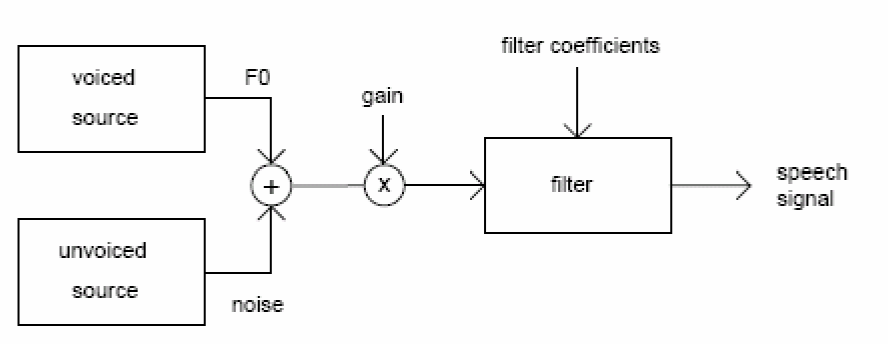

Fig. 3.1. Source-filter model of speech [1].

The formant filters represent only the resonances of the vocal tract, so additional provision is needed for the effects of the shape of the glottal waveform and the radiation characteristics of the mouth. There are two basic structures in general, parallel and cascade, but for better performance some kind of combination of these is usually used.

In formant synthesis, the set of rules controlling the formant frequencies and amplitudes and the characteristics of the excitation source is large. Also nasalized sounds are problematic with this approach.

### 3.3. Concatenative synthesis

One of the easiest ways of producing intelligible and natural sounding synthetic speech is the concatenation of the prerecorded natural speech samples. However such kind of synthesizers will be limited to one speaker and one voice.

The important aspects of concatenative synthesis are:

- Selection of correct unit length,
- Number of concatenation points and,
- Memory requirements.

The selection of unit is usually a trade-off between longer and shorter units. Selection of longer units results in more naturalness, less concatenation points and, more memory [1]. Shorter units require less memory but, more concatenation points are involved. The units usually used are: words, syllables, phonemes, diphones and, triphones.

### *3.3.1Unit selection*

*Demisyllables*- These are the initial and final parts of syllables. Using demisyllables, instead of phonemes and diphones, lesser number of concatenation points is required. One advantage of demisyllables is that only 1000 are required to construct about 10,000 syllables of English. Demisyllables take into account the co-articulation effects and the allophonic variations are also included. But, a pure demisyllable is not capable of producing all the sounds [1].

*Phonemes*- Phonemes are the most commonly used units in speech synthesis. The inventory of basic units is usually between 40 and 50, which is smallest compared to other units. However articulation effects must be individually accounted [1].

*Diphones*- Diphones are two phones followed one after another. There is a steady transition between the adjacent phones, which reduces the distortion due to concatenation points. They don't require co-articulation effects to be separately accounted. But, as the number of diphones in principle is square the number of phonemes they occupy more memory [1]. Longer segmental units, such as *triphones* or *tetraphones*, are quite rarely used.

Some problems in concatenative synthesis compared to other methods are:

- Distortion at the points of concatenation.
- High memory requirements.
- Gathering the samples from natural speech is time consuming.

### *3.3.2 Prosody manipulation*

One commonly used technique for synthesis of speech waveforms in TTS is concatenation of short speech units taken from a prerecorded inventory. After concatenation, these units are modified in duration and melody to smoothly join each other and achieve the prosody of a natural utterance. In order to perform these modifications without introducing artifacts, R.J. McAulay and T.F. Quatieri introduced sinusoidal model, based on which speech transformations can be performed. The advantage of this model is that it is capable of performing high quality time-scale and pitch-scale modifications of both speech and music signals. The next section discusses the features and various aspects of speech synthesis using sinusoidal model.

## 4. SPEECH ANALYSIS / SYNTHESIS: SINUSOIDAL REPRESENTATION

One of the possible representations of speech waveform can be viewing speech as the result of passing a glottal excitation waveform through a time-varying linear filter that models the resonant characteristics of the vocal tract [5]. It is appropriate to assume that the excitation signal, in most of the cases, be in one of the two possible states, corresponding to voiced or unvoiced speech. There are many approaches to model the glottal excitation [2]. In the sinusoidal model, the glottal excitation is represented in terms of sum of sine waves, which when applied to a time-varying filter, leads to the desired sinusoidal representation of the speech waveforms. This model was proposed by Robert J. McAulay and Thomas F.Quatieri. Hence this model is also known as McAulay/Quatieri model. In this section, the *sinusoidal model* for the speech waveform is discussed, which is characterized by amplitudes, frequencies, and phases of the component sine waves.

### 4.1 Sinusoidal Speech model

In the speech production model, the speech waveform *s(t)* is assumed to be the result of passing the glottal wave, *e(t)* through a time-varying vocal tract filter. If the time-varying impulse response of the vocal tract filter is *h(τ ; t),* then

$$s(t) = \int_0^t h(t-\tau)e(\tau)d\tau \,. \qquad (1)$$

In sinusoidal model the glottal excitation is represented in terms of sum of sine waves, which, when applied to the time-varying vocal tract filter, leads to the desired sinusoidal representation for speech waveforms. Thus the glottal excitation can be represented as

$$e(t) = \mathrm{Re}\sum_{l=1}^{L(t)} a_l(t)\exp\{j[\int_0^t \omega_l(\sigma)d\sigma + \varphi_l]\} \qquad (2)$$

where, for the $l^{\text{th}}$ sinusoidal component, $a_l(t)$ and $\omega_l(t)$ represent the amplitude and frequency and $\varphi_l$ represents a fixed offset which accounts for the fact that the sine waves will generally not be in phase [2]. Letting

$$H(\omega;\, t) = M(\omega;\, t)\, exp[j\,\zeta\, (\omega\, ;\, t)] \qquad (3)$$

represent the time-varying vocal tract transfer function, and assuming that the glottal excitation parameters in (2) are constant over the duration of impulse response of the vocal tract filter in effect at time *t*, then using (2 ) and (3) in (1) results in the speech model

$$s(t) = \sum_{l=1}^{L(t)} a_l(t) M[\omega_l(t);t].\exp\{j[\int_0^t \omega_l(\sigma)d\sigma + \zeta[(\omega_l(t);t] + \varphi_l]\}\,. \quad (4)$$

This can be more concisely written as

$$s(t) = \sum_{l=1}^{L(t)} A_l(t)\exp[j\psi_l(t)] \quad (5)$$

where

$$\begin{aligned} A_l(t) &= a_l(t)M[\omega_l(t);t] \\ \psi_l(t) &= \Omega_l(t) + \zeta\ [\omega_l(t),t] \end{aligned} \quad (6a,6b)$$

$$\Omega_l(t) = \int_0^t \omega_l(\sigma)d\sigma + \varphi_l \quad (6c).$$

Equations (6a) and (6b) represent the amplitude and phase of the $l^{th}$ sine wave along the frequency track $\omega_l(t)$. The equations (5) and (6) represent the speech wave as a sum of sine waves of arbitrary amplitudes, frequencies and phases.

### 4.2 Estimation of speech parameters using Sinusoidal analysis:

The speech parameters are estimated by applying short-time Fourier transform (STFT) to a quasi-stationary part of the speech waveform. The STFT of speech will have peaks occurring at all pitch harmonics. Therefore the frequencies of underlying sine waves correspond to the peaks of STFT. The amplitudes and phases of the component waves are estimated at the peaks from the high resolution STFT using a simple peak-picking algorithm [2]. The basic analysis system for the sine model is shown in Figure 4.1.

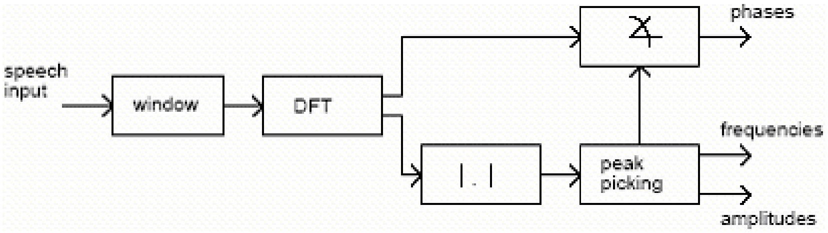

Fig. 4.1. Sinusoidal Analysis system [2].

The problem in analysis/synthesis is to take a speech waveform, extract the parameters that represent a quasi-stationary portion of that waveform, and use those parameters or coded versions of them to reconstruct an approximation that is as close as possible to the original speech [2]. The approach proposed by Robert J. McAulay and Thomas F.Quatieri is that an estimator will be derived based on a set of idealized assumptions; then, once the structure of the ideal estimator is known, modifications will be made as the assumptions are relaxed to better model practical speech waveforms [2].

In the first step, the time line is broken down into contiguous sequence of frames, each of duration T. The centre of the analysis window for the $k^{th}$ frame occurs at time $t_{k.}$ Assuming that the vocal tract and glottal parameters are constant over an interval of time that includes the duration of the analysis window and the duration of the vocal tract response, then (6b) can be written as

$$\psi_l(t) = \omega_l^k(t - t_k) + \theta_l^k, \qquad (7)$$

where the superscript "*k*" is used to indicate that the parameters of the model may vary from frame to frame. As a consequence of (7) the synthetic speech waveform for the $k^{th}$ frame is given by

$$s(n) = \sum_{l=1}^{L^k} \gamma_l^k \exp(jn\omega_l^k) \qquad (8)$$

where $\gamma_l^k = A_l^k \exp(j\theta_l^k)$ represents the complex amplitude for the $l^{th}$ component of the $L^k$ sine waves. Now the problem is to find the parameters of *s(n)* such that the synthetic speech waveform in (8) fits to the measured waveform denoted by *y(n)*. A useful criterion for judging the goodness of the fit is to find the minimum mean-squared error [2],

$$\varepsilon^k = \sum_n |\, y(n) - s(n)\,|^2 \ . \qquad (9)$$

It can be proved that the error is minimized by selecting all the harmonic frequencies in the speech bandwidth.

*Assumptions-* The above result is arrived at based on the following assumptions:

1. The input signals are perfectly voiced speech,

2. Pitch period is assumed to be constant over the duration of $k^{th}$ frame,
3. The pitch period is known and the width of analysis window is a multiple of the pitch period.

*Parameter estimation for unvoiced speech-* The same estimator structure can be applied even when the ideal voiced speech assumption is no longer valid [2]. The same estimator can be applied for the case of unvoiced speech provided the analysis window is “wide enough” that to satisfy

$$|\omega_i^k - \omega_l^k| \geq \frac{4\Pi}{N+1}. \qquad (10)$$

Since the spectral changes for the unvoiced speech are rapid the spectral width of the window should be less so as to track all the rapid changes. For example if the peaks in periodogram are 100 Hz apart, then the window width is constrained to be at least 20ms wide [2].

### 4.3 The Sinusoidal synthesis system

The sinusoidal analysis as explained in the previous section is done for continuous frames of speech waveform. This results in representation of speech waveform in terms of amplitudes, frequencies, and phases of a set of sine waves. But as these parameters are time varying in nature, the straightforward method of summing the sine waves of different parameters results in discontinuities at the frame boundaries, which seriously degrades the quality of synthetic speech. Therefore the set of parameters obtained over one frame should be matched with those of the consecutive frame [3]. To achieve this frequency tracking is done. Figure 4.2 is the basic block diagram for sinusoidal speech synthesis proposed by R. J. McAulay and T. F. Quatieri.

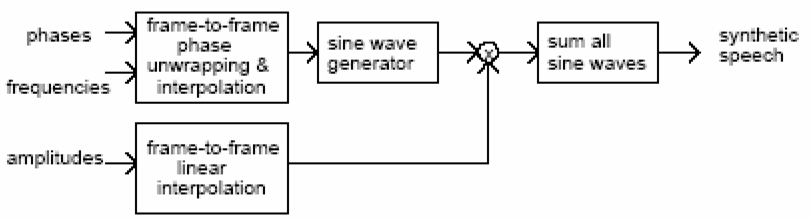
phases
frequencies
frame-to-frame phase unwrapping & interpolation
sine wave generator
sum all sine waves
synthetic speech
amplitudes
frame-to-frame linear interpolation

Fig. 4.2. Sinusoidal synthesis system [2].

In this section the frequency tracking algorithm and the interpolation technique are discussed.

### *4.3.1 Frame-to-Frame peak matching:*

Peak matching would be straightforward if number of peaks over the frames remained constant [3]. However, the number of peaks estimated on one frame will not be the same in other frames because of one of the following factors [2]:

1. Spurious peaks that can come and go due to side lobe interactions,
2. Changes in the pitch implying changes in location of peaks,
3. Rapidly changing regions of speech, voiced or unvoiced transitions.

So, in order to account for such rapid movements in spectral peaks, the concept of "birth" and "death" is used [3].

*The matching algorithm:*

*Step 1:* The frame (i.e., frame $k$ or $k+1$) with the greater energy is chosen as the "anchor" frame. This is the frame for which there is the most confidence in the chosen peaks. Assume (without loss of generality) that frame $k$ is the anchor frame.

*Step 2:* An attempt is made to find a frequency match $g_m$ from frame $k+l$ for each of the chosen frequencies of frame $k$ (where p is the number of estimated peaks). Specifically, $f_n$

of frame $k$ matches to $g_m$ of frame $k+1$, if

$$| f_n - g_m | < | f_n - g_i | < \delta$$

i.e., if $g_m$ falls within distance $\delta$ of $f_n$ and $g_m$ is the closest such frequency. When a match is made, then a candidate match is obtained and Step *3* is executed. If no match is made, it is assumed that the frequency n has "died" on entering frame $k+1$, so that n matches to itself with zero magnitude. If a match is not made, Step 2 is repeated with the next frequency, $f_{n+1.}$

*Step 3:* In this step, candidate matches from Step 2 are verified. Suppose that a frequency $f_n$ of frame $k$ has been tentatively matched to frequency $g_m$ of frame $k+1$. Then, if $g_m$ has no better adjacent frequency match within frame $k$, the candidate match is a true match. More specifically, the following conditions must hold:

$$| g_m - f_n | < | g_m - f_{n-1} | \quad \text{and}$$

$$| g_m - f_n | < | g_m - f_{n+1} |$$

If these conditions are violated, it is concluded that the frequency *n* "dies" on entering frame *k+1.* Consequently, the frequency $f_n$ matches to itself with zero magnitude. Finally, the algorithm returns to Step 2 to search for a match to the next frequency $f_{n+1}$.

*Step 4:* When all frequencies of frame *k* have been matched to some frequency in frame *k+1*, frequencies in frame *k+l* for which no match was found are determined. Suppose that $\mathrm{g_i}$ is one such frequency. Then $g_i$ was born in frame *k* and thus its match is created by specifying a new frequency $g_i$, in frame *k* with zero magnitude. This is done for all such unmatched frequencies. The various cases of the matching procedure are illustrated in figure 4.3.

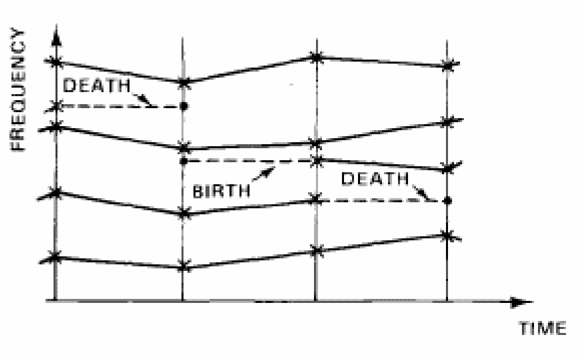

Fig. 4.3. Nearest neighbor matching with birth-death mechanism [3].

*4.3.2 Direct Interpolation technique:*

The basic underlying principle of this method is that frequency, amplitude and phase terms in the equation are made to vary for each sample inside the frame. There is a smooth transition of parameter values from beginning frame boundary to the end frame boundary [2].

Let $(A_l^k,\omega^k{}_l,\theta^k{}_l)$ and $(A_l^{k+1},\omega^{k+1}{}_l,\theta^{k+1}{}_l)$ denote the parameters of two successive frames for the $l^{th}$ harmonic. For amplitude interpolation, a very simple expression is used:

$$A(n) = A^k + \frac{(A^{k+1} - A^k)}{S} n, \qquad (11)$$

But this system cannot be used to interpolate the frequency and phase because the measured phase is obtained modulo 2Π. Hence, phase unwrapping must be performed to ensure that the harmonics are maximally smooth across frame boundaries. The first step in solving this problem is to postulate a phase interpolation function that is a cubic polynomial, namely,

$$\bar{\theta}(t) = \zeta + \gamma t + \alpha t^2 + \beta t^3. \qquad (12)$$

Since this method requires computing the parameter set for each sample of the frame, it is highly computation extensive. Therefore if a particular application requires a higher frame rate, this method may not be suitable.

### 4.4 Magnitude-only reconstruction using Sinusoidal model:

R. J. McAulay and T. F. Quatieri, proposed a sinusoidal model for the speech waveform that requires specification of only the amplitudes and frequencies of the component sine waves [3]. These parameters are estimated from the short-time spectral magnitude. The resulting synthetic waveform preserves the short-time spectral magnitude during rapid movements of spectral energy such as voiced/unvoiced transitions, and yields speech of very high quality and intelligibility. One of the classical results of speech perception is the axiom that the ear is sensitive principally to the short-time spectral magnitude and not the phase, provided phase continuity is maintained [3].

If s(n) represents the sampled speech waveform then

$$s(n) = \Sigma\, a_i(n) \sin[\, \Phi_i(n)\, ] \qquad (13)$$

where $a_i(n)$ and $\Phi_i(n)$ are the time-varying amplitudes and phases of the i'th tone. The phase is taken to be the integral of the instantaneous frequency $f_i(n)$ and therefore satisfies the recursion

$$\Phi_i(n) = \Phi i(n-1) + 2\Pi f_i(n)/f_s \qquad (14)$$

where $f_s$ is the sampling frequency. If the tones are harmonically related then $f_i(n) = i{*}f_0(n)$, where $f_0(n)$ represents the fundamental frequency at time n. One particularly attractive property of the above model [3] is the fact that phase continuity, hence waveform continuity, is guaranteed as a consequence of the definition of phase in terms of the instantaneous frequency. This means that waveform reconstruction should be possible from the magnitude-only spectrum since a high-resolution spectral analysis should reveal the amplitudes and frequencies of the component sine waves. Of course the reconstructed phase function will differ from that of the original speech waveform, but this will be perceptually irrelevant provided we have successfully reconstructed the details of the Fourier transform magnitude.

### 4.5 Speech transformations using Sinusoidal model:

The following speech transformations can be done as desired using the sinusoidal model [4].

- Time-scale modification: the rate of articulation can be slowed down so as to make the speech comprehensible or speech can also be compressed so as to fit into a time-interval,
- Frequency scaling: bandwidth reduction where frequency is compressed,
- Pitch scaling: in psycho acoustic research.

To perform these transformations, the excitation and vocal tract system parameters are extracted and scaled accordingly. The extraction of these parameters is illustrated in Figure 4.4. The excitation frequencies are estimated by picking the peaks of the uniformly spaced samples of the short-time Fourier transform magnitude. The system amplitude and phase are estimated by homomorphic deconvolution. Finally the excitation amplitude is obtained by dividing the measured amplitude by the system amplitude, and the excitation phase from the system phase.

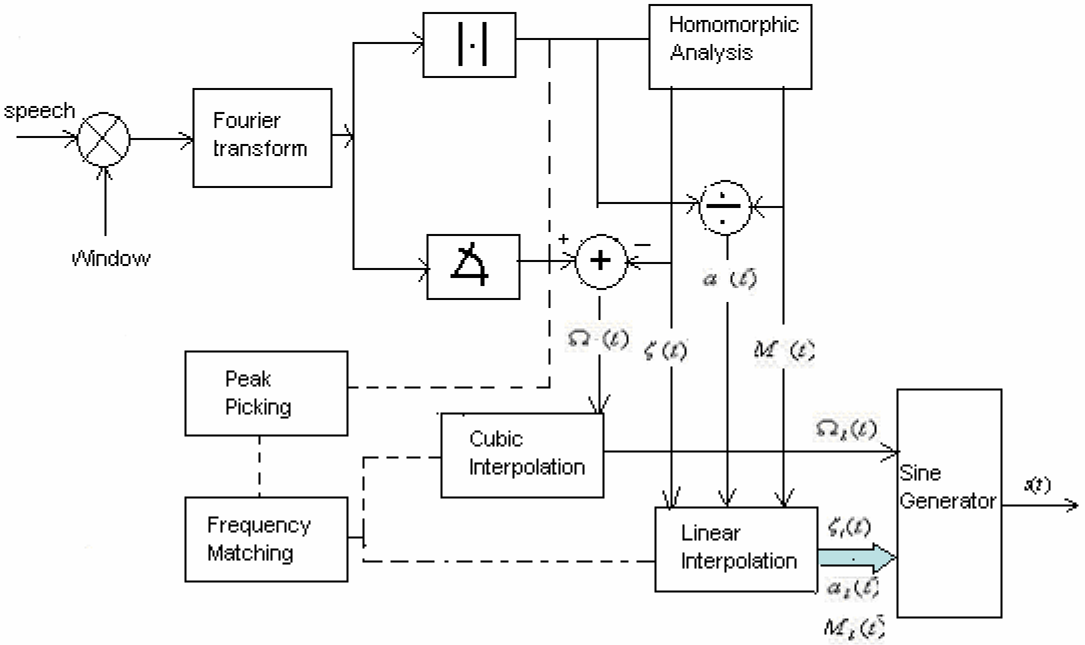

Fig. 4.4. Sinusoidal analysis/synthesis system for parameter extraction [4].

This section briefly discusses the basic idea in achieving these transformations based on sinusoidal model.

### *4.5.1 Time-scale modification:*

In time-scale modification the goal is to achieve the change in the *rate of articulation* while maintaining the perceptual quality of the original speech unchanged. This implies that the frequency trajectories of the excitation (and hence the pitch contour) are stretched or compressed in time and the vocal tract changes at a slower or faster rate. The sinusoidal model is ideal to achieve this, since it provides the flexibility to control the required parameters.

For a fixed rate change ρ, the time-warping transformation, denoted by W(t), is given by

$$t_0 = W(t_0^{\sim}) = \rho\, t_0^{\sim}$$

where $t_0^{\sim}$ the time at the original rate and $t_o$ is the transformed time. Thus speech events take place at a time $t_0$ according to the new time scale will have occurred at $\rho^{-1}t_0$ in the original time scale. When the sine model is applied the events which are time scaled are the system amplitudes and phases, and the excitation amplitudes and frequencies, along each frequency track. Since the parameters of the unmodified synthesis are available as continuous functions of time, then theoretically, any rate change is possible [4]. Time-varying rate change is also possible. Here the time-warping transformation is given by

$$t_0 = W(t_0^{\sim}) = \int_0^{t0\sim} \rho(\tau) d\tau\,,$$

where $\rho(\tau)$ is the desired time-varying rate change. Speech events will take place at a time $t_0$ in the new time scale will now occur at a time $t_0^{\sim} = W^{1}(t_0)$ in the original time scale.

### *4.5.2 Frequency transformations:*

Since the synthesis procedure consists of adding up the sinusoidal waveforms for each of the measured frequencies, the procedure is ideally suited for performing various frequency transformations.

*Spectral Warping:*

This is achieved by scaling each of the frequency tracks $\omega_l(t)$ by a desired factor $\beta$ which results in the scaled excitation phase. The original amplitudes will get shifted to the new scaled frequency track locations.

*Pitch Modification:*

In pitch modification, the excitation function is frequency scaled as above, while the spectral envelope of the system function, and thus of the speech waveform, is unchanged. The result is to modify the fundamental frequency of the speaker. In the context of the sinusoidal analysis/synthesis system, the frequency of each sine-wave component is scaled by a desired factor to generate a new frequency track. This results in a scaled excitation phase. The original excitation amplitude is simply shifted to the new frequency track locations. To preserve the short-time spectral envelope, the system amplitudes and phases are found at the new frequency track locations.

**4.6 Prosody modification using sinusoidal model:**

The three main prosodic or suprasegmental features of spoken language are: pitch, duration, and stress over the time [1]. Good control of prosody and intonation, results in good quality synthetic speech. However, the prosody of speech also depends upon gender, age, emotions, and other characteristics of the speaker. This multivariable dependence of prosody makes accurate modeling very difficult. Also the prosodic features are not independent, which further aggravates the problem. But, using the sinusoidal transformations it is possible to achieve desired prosody into the speech. Since the sinusoidal model is capable of performing speech transformations like, pitch scaling, time scaling and frequency scaling both individually and simultaneously, it can be a solution to the problems in prosody. Such a possible scheme for achieving prosody and hence expression into synthetic speech is illustrated in Figure 4.5. Where $\Omega_l(t), \zeta_l(t), a_l(t), M_l(t)$ are the parameters obtained from sinusoidal analysis (shown in Fig.4.4.), and $\Omega_l^{\sim}(t), \zeta_l^{\sim}(t), a_l^{\sim}(t), M_l^{\sim}(t)$ are the parameters obtained after prosody modification. These modified parameters are resynthesized using the sinusoidal model to get modified speech waveforms.

*Prosodic features:*

The *pitch pattern* over a sentence (intonation) in natural speech is a combination of many factors. The pitch contour depends on the meaning of the sentence. It is also affected by gender, physical and emotional state, and attitude of the speaker. For example, in normal speech the pitch slightly decreases toward the end of the sentence and when the sentence is in a question form, the pitch pattern will raise to the end of sentence. Pitch scaling can be done as explained in 4.5.2.

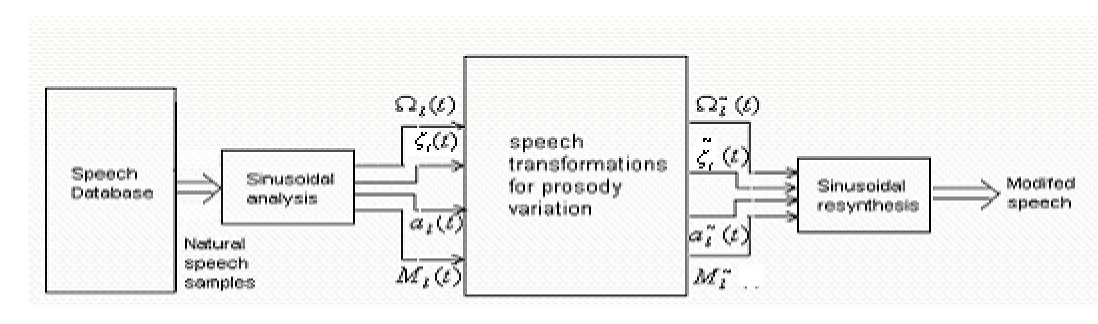

Fig. 4.5 Scheme for changing Prosody.

The *duration or time* characteristics of the speech are dependent upon speaking rate, and rhythm. For example, consonants in non-word-initial position are shortened, emphasized words are significantly lengthened, or a stressed vowel or sonorant preceded by a voiceless plosive is lengthened [1]. This can be easily achieved by time-scale modification as explained in 4.5.1.

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

The speaker's feelings and emotional state affect speech in many ways. The proper implementation of these features in synthesized speech will 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 [1].

## 5. DISCUSSION

Implementation of prosodic features in synthesized speech is probably one of the most challenging tasks in speech synthesis. In this report, an attempt is made to study sinusoidal analysis and synthesis system for applying the model in achieving prosody in the synthesized speech. However, prosody is not a function of single variable and involves several aspects such as state of the speaker, context, gender, age, etc. Almost everything seems to have effect on prosodic features of natural speech which makes accurate modeling very difficult. Since, the method of sinusoidal representation allows very flexible control over several speech parameters such as excitation frequency, amplitude and phase, it is possible to achieve desired prosody modification in the synthetic speech.

**Acknowledgement**

I hereby sincerely express my gratitude to Professor Preeti S. Rao for her timely advices and constant support in preparing this seminar report. Her encouragement and clarifications helped me to gain insight into the subject.

# References

[1] Sami Lemmetty, "Review of Speech Synthesis Technology," *Master's Thesis*, Dept. of Electrical and Communication Engineering, Helsinki University of Technology, March 30, 1999.

[2] R.J. McAulay and T.F. Quatieri, "Speech analysis-synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, no. 4, pp. 744-754 August 1986.

[3] R.J. McAulay and T.F. Quatieri, "Magnitude-only reconstruction using a sinusoidal speech model," in *Proc. Int. conf. Acoust., Speech, Signal Preocessing,* San Diego, CA, 1984, p.27.6.1.

[4] T.F. Quateiri and R.J. McAulay, "Speech transformations based on a sinusoidal representation," in *Proc. Int. Conf. Acoust., Speech, Signal Processing,* Tampa, FL, 1985, p.489.

[5] O'Saughnessy D. (2001). *Speech Communications – Human and Machine*, University Press.