M.Tech. Credit Seminar Report, Electronic Systems Group, EE Dept, IIT Bombay, submitted November '04

SPEECH ANALYSIS-SYNTHESIS FOR SPEAKER CHARACTERISTIC MODIFICATION

G. Gidda Reddy (Roll no. 04307046)

(Supervisor: Prof. P. C. Pandey)

ABSTRACT

Speech analysis-synthesis techniques, for speaker characteristic modification, have many applications. Such a modification of speech could potentially be helpful to people with hearing disabilities. This paper describes the HNM (Harmonic plus Noise Model) [1], a new analysis/modification/synthesis model, in which each segment of speech can be modeled as two bands: a lower "harmonic" part can be represented using the amplitudes and phases of the harmonics of a fundamental and an upper "noise" part using an all pole filter excited by random white noise, with dynamically varying band boundary. HNM based synthesis [2] can be used for good quality output with relatively small number of parameters. Using HNM, pitch and time scaling are also possible without explicit estimation of vocal tract parameters. This paper also describes some other analysis-synthesis techniques (LPC vocoders, Cepstral vocoders, and Sine Transform Coders) in brief.

1. INTRODUCTION

High quality speech modification is a subject of considerable importance in many applications such as text to speech synthesis based on acoustic unit concatenation, psychoacoustics experiments, foreign language learning, etc.

Speech modification techniques have many applications [5]. Speeding up a voice response system save time for a busy, impatient user. Compressing the spectrum could be of helpful to people with hearing disabilities. Deep-sea divers speak in a helium-rich environment, and this gives the speech an effect called Mickey Mouse effect, which is due to the spectral changes caused by changes in the velocity of sound [5], where the spectral modification techniques are useful. A popular game is to change a male voice into female voice and vice versa, such a game could conceivably be part of a psychological gender experiment. Voice modification techniques may also be applicable to Automatic Speech Recognition tasks [5]. One more application of speech modification is in a long distance communication link, in which atmospheric conditions result in occasional fading of the signal [5].

Standard techniques for speech modification include LPC vocoders [6], and digital phase vocoders [5]. In LP based methods, modifications of the LP residual have to be coupled with appropriate modifications of the vocal tract filter. If the interaction of the excitation signal and the vocal tract filter is not taken into account, the modified signal will be degraded. This interaction seems to play a more dominant role in speakers with high pitch (as in the case of females and children voices). This is a possible reason for the failure of LPC based methods. Digital phase vocoder is computationally intensive and often generates reverberation.

In the past few years a number of alternative techniques have been proposed including time domain and frequency domain PSOLA and different methods developed around the sinusoidal models of Quetieri and McAulay. PSOLA synthesis [7] scheme allows high quality pitch and time scale transformations, at least for moderate values of modification parameter. Because PSOLA synthesis scheme is a nonparametric model (which assumes no specific model for the speech signal, except that it is locally periodic on voiced portions), it does not allow complex modifications of the signal such as increasing the degree of friction, or changing the amplitude and phase relationships between the pitch harmonics.

Since the sinusoidal model [8] assumes the representation of speech waveform as a summation of finite number of sinusoids with arbitrary amplitudes, frequencies and phases, it does not allow complex modifications, which suffers from same drawbacks as given above for PSOLA.

This report describes a flexible synthesis model called 'HNS' (Harmonic plus Noise Synthesis) [1], based on a harmonic plus noise decomposition where the harmonic part accounts for the periodic structure of the speech signal and the noise part accounts for the non-periodic structure of the speech signal such as fricative noise, period to period variation of the glottal excitation. HNM has a capability of providing high quality speech synthesis and prosodic modifications.

One main drawback of this model (HNM) is its complexity. This report describes four methods of reducing the complexity of HNM [4], which includes Straight Forward synthesis (SF), synthesis using Inverse Fast Fourier Transform method, synthesis using recurrence relations for trigonometric functions (RR), and synthesis using Delayed Multi-Resampled Cosine functions (DMRC). Higher quality of speech synthesis can be obtained by the HNM using DMRC method when compared with the other methods.

2. SPEECH SYNTHESIS TECHNIQUES

2.1. History of speech synthesis

Many years ago (in 1791) Von Kempelen demonstrated that the speech production system of the human being could be modeled. He showed this by building a mechanical [5] contrivance that talked. Wheatstone built a speaking machine [5] that was based on Von Kempelens work. Much later, Riesz built a mechanical speaking machine [5] that was more precisely modeled after the human speech producing mechanism.

Homer Dudley pioneered the development of channel vocoder (voice coder) and the voder (voice operated demonstrator) [5]. It is important to realize that voder did not speak without a great deal of help from a human being. This difficulty was eliminated in Haskins Play Back. Many speech-synthesis devices [5] were built in the decades following the invention of the voder, but the underlying principle has remained quite fixed as that of the voder controls.

2.2. Speech synthesis techniques

The use of talking machines provides flexibility for extended or even arbitrary vocabularies, which are required for applications such as unlimited translation from written text to speech. The basic approaches for sound unit to speech translation are:

A. Articulatory synthesis

B. Source filter synthesis (synthesis by rule)

C. Concatenative synthesis

2.2.1. Articulatory synthesis

The articulatory synthesis model consists of physical models for their articulators and their movements. The mechanical systems built by Von Kempelen and Wheatstone [5] are belonging to this category. Now in the approaches, described by Coker and Colleagues, computational models for the physical systems are used directly to estimate the resulting speech signal. This method is appealing because of its directness, but the difficulty of deriving the physical parameters by analysis and the large computational resources required for synthesis have made this approach more interesting from a scientific point than a practical one.

2.2.2 Source-Filter synthesis

Source-Filter synthesis [5], also called formant synthesis, is based on spectral shaping of driving excitation, which has most often used formants to characterize the spectral shape. Formants have a straight forward acoustic phonetic interpretation, while being computationally simple compared to full articulatory models. A formant is usually represented by a second order filter. This type of synthesis is more applicable in case of parameter modifications for a particular context.

Various configurations of Formant synthesizers are:

A. Formant analysis synthesizers

B. LPC analysis synthesizers

C. Cepstral analysis synthesizers

D. Channel vocoders

E. Parallel formant synthesizers

This paper describes a high quality synthesizer using HNM (Harmonic plus Noise Model) and permits the description of these models.

2.2.3. Concatenative synthesis

These approaches have been used by many systems, in which speech waveforms are stored and then concatenated during synthesis. In most of the cases voiced sounds will be compressed by manipulating a pitch period waveform to reduce the number of signal samples that are required to have a power spectrum that should be sufficiently close to the original.

A technique that has been used by many systems is Pitch Synchronous Over-Lap and Add method (PSOLA), in which diphones are concatenated pitch synchronously. The alignment to pitch periods permits variation of pitch frequency by varying the timing of repeats for each waveform types. There are different PSOLA [7] techniques have been developed. The above one is called TD-PSOLA. Other techniques include frequency domain PSOLA, LP-PSOLA (Linear Predictive PSOLA, in which linear prediction coefficients will be stored to represent a segment rather than storing the diphone waveforms), MPLPC-TDPSOLA (Multi Pulse Linear Predictive Coding PSOLA, in which the input is a multi pulse sequence of pulses), RELP-

PSOLA (Residual-Excited Linear Predictive PSOLA), and Code-Excited PSOLA (CEPSOLA). Sinusoidal modeling of speech [8] is also come in to this category, in which the storage of variables representing the segments is that of sinusoidal signals.

3. SPEECH TRANSFORMATIONS USING ANALYSIS-SYNTHESIS SYSTEMS

3.1. LPC vocoders

Vocoders are analysis-synthesis systems. So, once the parameters of a given speech signal are analyzed, it is possible to intervene before synthesis to produce some transformed version of speech. Linear Predictive Analysis [6] is a powerful tool by which speech can be synthesized and transformed suitably into the required forms. LPC analysis assumes an all-pole model, which is represented as,

$$H(z) = \frac{1}{1 - \sum_{k=1}^{P} a_k z^{-k}} \tag{1}$$

where $P=2*(B+1)$, and B represents the number of formants with in the bandwidth, and a_k, for $k=1....P$, are the coefficients of the P-th order polynomial. From the above equation the discrete time response $y(n)$ of the system to an excitation signal $x(n)$ is given by,

$$y(n)=x(n)+\sum_{k=1}^{P} a_k\, y(n-k) \tag{2}$$

The coefficients for the second term of this expression are generally computed to give an approximation to the original sequence, which will yield a spectrum for $H(z)$ that is an approximation to the original speech spectrum. Thus, here the prediction is the speech signal by a weighted sum of its previous values given by,

$$y'(n)= \sum_{k=1}^{P} a_k\, y(n-k) \tag{3}$$

This has the form of FIR filter, but when it is included in the previous expression the resulting production model is IIR. The coefficients that yield the best approximation of $y'(n)$ to $y(n)$, in the mean square sense, are called the Linear Prediction Coefficients. In the statistical literature the overall model is sometimes called as AR (Auto-Regressive) model. The difference between the predictor and the original signal is referred to as the error signal (also called Residual error) given by $e(n)=y(n)-y'(n)$. When the coefficients are chosen to minimize the error signal energy, the resulting error signal can be viewed as an approximation to the excitation function. The prediction error has large peaks that occur once per pitch period.

3.1.1. Time scale modifications using LPC vocoder

It is assumed that, during synthesis, the number of samples synthesized will be made equal to the number of samples analyzed. If it is made such that the number of samples synthesized is different from the number of samples analyzed, which effectively changes the duration of output speech relative to input speech, then the resultant speech represents time-scale

modification one. The fundamental frequency and the spectral parameters have been unchanged.

3.1.2. Spectral modifications using LPC vocoder

In an LPC vocoder spectral modifications can be implemented in various ways [6]. For example, once the analyzer has determined the synthesizer parameters, the spectral envelope can be computed, either directly or by computing the DFT of the synthesizer impulse response. A new set of auto-correlation values are then computed from the modified spectrum and the reflection coefficients are recomputed. Alternatively, DFT of the computed correlation values yields the square of the spectral magnitude, which can now be modified and an inverse DFT computed to create the modified correlation function, which can then be used to compute the modified parameters for transmission.

3.1.3. Pitch scale modifications using LPC vocoder

Using an LPC technique it is possible to estimate the spectral envelope (spectrum of vocal tract impulse response), which can be used to create a time domain inverse filter. By passing the original speech signal through an inverse filter we can get an approximation to the excitation. By low-pass filtering and sampling at the required rate (that is modifying the excitation) and then convolving it with vocal tract impulse response we will get the pitch modified speech.

The main drawback of an LPC analysis is, in LP based methods, modifications of the LP residual have to be coupled with appropriate modifications of the vocal tract filter. If the interaction of the excitation signal and the vocal tract filter is not taken into account, the modified signal will be degraded. This interaction seems to play a more dominant role in speakers with high pitch. And also this model (all pole model) is not suitable for all phonemes (such as nasal sounds).

3.2. Cepstral vocoders

This scheme (Cepstral vocoder) was developed by Oppenheim (in the late 1960s) which is a complete analysis-synthesis system based on homomorphic (that is Cepstral) processing [6]. We know that the spectrum of the speech signal can be represented as the product of the excitation spectrum and the vocal tract filter spectrum given by

$$|X(\omega)| = |E(\omega)||V(\omega)| \tag{4}$$

Taking the logarithm on both sides of the above equation, we will get

$$\log|X(\omega)| = \log|E(\omega)| + \log|V(\omega)| \tag{5}$$

From the above equation it is clear that the logarithmic spectrum is separated as two parts namely, the log spectral components that vary rapidly with ω (high-time components, first term in the right side of the above equation) and the log spectral components that vary slowly with ω (low-time components, second term in the right side of the above equation). Hence using an appropriate filter we can separate the two components namely, the excitation spectrum and the vocal tract filter spectrum. This process is called deconvolution. The cepstrum is given by taking the inverse Fourier transform of the above equation given by

$$c(n)=\frac{1}{2\pi}\int_{-\pi}^{\pi}\log\left|X(\omega)\right|e^{i\omega n}d\omega \tag{6}$$

where $c(n)$ is the n-th cepstral coefficient. Thus the contribution of the excitation and the vocal tract filter can be separated in cepstral domain. Both components can be inverted to generate the original spectral magnitudes. The following block diagram describes the cepstral analysis method.

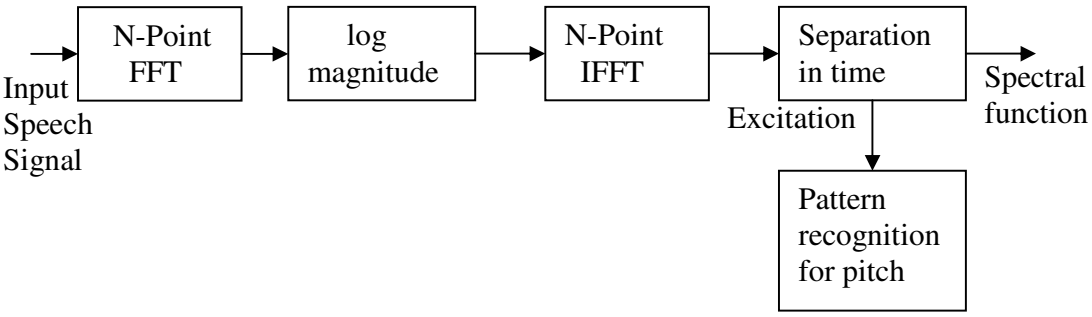

Fig.1. Description of Cepstral analysis from [5]

3.2.1. Timescale scale modifications using Cepstral vocoder

One method of performing time-scale modification is to alter both the fundamental frequency parameters and the spectral parameters and then to modify the ratio of the input to output sampling rates. Comparable manipulations will allow for time scale modifications in LPC vocoders also. One more method of performing time-scale modification is, if it is assumed that, during synthesis, the number of samples synthesized will be made equal to the number of samples analyzed. If it is made such that the number of samples synthesized is different from the number of samples analyzed, which effectively changes the duration of output speech relative to input speech, then the resultant speech represents time-scale modification one. The fundamental frequency and the spectral parameters have been unchanged.

3.2.2. Spectral modifications using Cepstral vocoder

The following diagram describes the spectral modifications using the cepstral analysis-synthesis method. By passing the cepstrum through a low time lifter and then applying DFT, the logarithm spectral envelope will be generated. By taking exponentiation the spectral envelope will be generated. Then the spectral envelope will be modified in a desired manner and then applying inverse DFT the modified vocal tract impulse response will be generated.

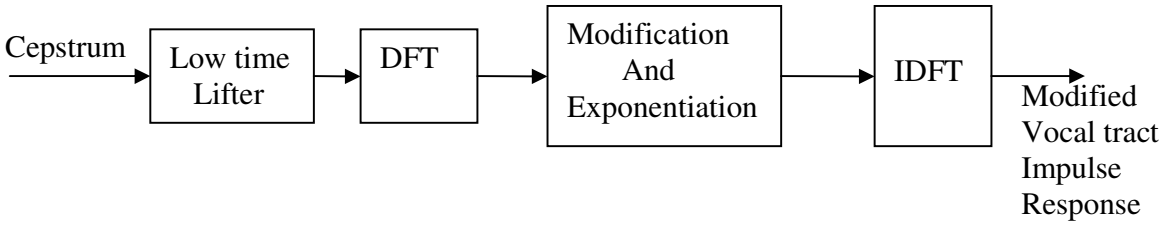

Fig.2. Spectral modifications using Cepstral vocoder from [5]

3.2.3. Pitch scale modifications using Cepstral vocoder

Using Cepstral analysis we can separate the excitation function and the vocal tract impulse response. From an excitation function we can extract the pitch. This will be modified in a desired manner and will be convolved with the vocal tract impulse response (which can be obtained by low time liftering the Cepstral function and then taking the DFT and then taking exponentiation and IDFT) to get the desired pitch scale modification of speech.

The main drawback of this system is its complexity when compared to other methods due to the calculations of DFTs and IDFTs are involved. And also this method has the difficulty with liftering, which introduces errors while separating the excitation function and vocal tract impulse response function.

3.3. Sine transform coder (STC)

Quatiery and McAulay developed STC (Sine Transform Coder) [8]. In this method the synthesizer is excited by a collection of sinusoidal signals. The frequencies and magnitudes of these signals are derived with an analysis procedure based on a high resolution, short-time DFT. The sum of these sinusoids represents the resultant synthesis.

3.3.1. Time scale modifications using STC

Time scale modifications using this method [9] are as follows. The analysis procedure computes the frequencies and magnitudes of the sinusoids at a rate corresponding to the rate at which successive DFT s is performed. When the rate of presentation of these parameters to the synthesizer is changed, the rate of resultant synthetic speech is also changed.

3.3.2. Spectral modifications using STC

From the given high-resolution DFT we can find the spectral envelope using different schemes such as LPC analysis, Cepstral analysis. By scrunching the spectral envelope, new magnitudes [9] will be assigned to the sinusoids based on sampling the scrunched spectrum, which gives the spectral modification.

3.3.3. Pitch scale modifications using STC

Pitch modification [9] can be done, from the given spectral envelope, by changing the derived frequencies and then sampling the spectral envelope at the new frequencies to generate new magnitudes for the shifted frequencies.

The drawback of STC model is, it assumes the representation of speech waveform as a summation of finite number of sinusoids with arbitrary amplitudes, frequencies and phases [8], and hence it does not allow complex modifications, such as increasing the degree of friction, or changing the amplitude and phase relationships between the pitch harmonics.

The other vocoders are channel and phase vocoders. The oldest form of speech coding device is the channel vocoder which was invented by Dudley. Another type of vocoder is the phase vocoder, which was originated and intensively investigated by Flanagan and Golden. The phase vocoder [5] begins by performing a spectral analysis of the incoming signal by means of FFT, which will produce the result as a real and imaginary component at each frequency position. By implementing the rectangular to polar co-ordinate transformation on the above result produces the result consists of magnitude and phase, which may be modified independently. The main drawback of phase vocoder is, phase vocoders are computationally

intensive and often generates reverberation. This paper mainly describes the HNM method and permits the description of these vocoders (channel and phase vocoders).

4. HARMONIC PLUS NOISE MODELS

HNM assumes the speech signal to be composed of two parts namely harmonic part and noise part [2]. The harmonic part accounts for the quasi-periodic components of the speech signal while the noise part accounts for the non-periodic components of the speech signal such as fricative or aspiration noise, period-to-period variations of the glottal excitation, etc. The two components are separated in the frequency domain by a time varying parameter, referred to as maximum voiced frequency, *Fm*. The lower band of the spectrum, below the maximum voiced frequency, is assumed to be solely represented by harmonics while the upper band, above the maximum voiced frequency, is represented by a modulated noise component. Even though these assumptions are not clearly valid from a speech production point of view they are useful from a perception point of view, they lead to a simple model for speech, which provides high quality synthesis and modifications of the speech signal.

Therefore the speech signal is represented as the sum of harmonic signal *h(t)* and the noise signal *n(t)*. Where *h(t)* is given by

$$h(t) = \sum_{k=1}^{K(t)} A_k(t)\cos(k\theta(t) + \phi_k(t)) \tag{7}$$

with $\theta(t) = \int_{-\infty}^{t} \omega_o(l)dl$. $A_k(t)$ and $\phi_k(t)$ are the amplitude and phase at time t of the k-th harmonic, $\omega_o(t)$ is the fundamental frequency and *K(t)* is the time varying number of harmonics included in the harmonic part.

The upper band which contains the noise part, is modeled by an AR model and is modulated by filtering a white Gaussian noise *b(t)* by a time varying normalized all-pole filter $h(\tau, t)$ and multiplying the result by an energy envelope function *w(t)*:

$$n(t)=w(t)[h(\tau, t)*b(t)] \tag{8}$$

4.1. Analysis of speech using HNM

The analysis scheme using HNM [3] is shown in figure. The analysis using HNM is based on frame-by-frame basis. The analysis consists of estimation of the parameters such as, whether the frame is voiced or unvoiced, if voiced, pitch, maximum voiced frequency *Fm* and the amplitudes and phases of harmonics of the fundamental frequency, if unvoiced, energy envelope and LPC coefficients (corresponding to the LPC all pole filter that are used for the noise part). The analysis and synthesis using HNM is pitch-synchronous and hence it is necessary to estimate the glottal closure instants (GCI s) precisely which can be estimated using the electroglotto-gram waveform from an impedence glottography.

Initially the speech signal should be applied to the voicing detector in order to detect whether the frame is voiced or unvoiced. Then for each voiced frame we have to estimate the maximum voiced frequency *Fm*. By analyzing the voiced frame at each glottal closure instant the amplitudes and the phases of all the pitch harmonics should be calculated up to the maximum voiced frequency *Fm*. From these parameters (pitch, maximum voiced frequency

(and hence the number of harmonics), and amplitudes and phases of all the pitch harmonics) it is possible to estimate the harmonic part of the HNM.

By subtracting the estimated harmonic part of the HNM from the original speech signal, we will get the noise part of the HNM (since HNM assumes that speech signal can be represented as the sum of the harmonic and noise parts). This noise part should be analyzed for estimating the LPC coefficients for a particular order and the energy envelope. The length of the analysis window for noise part will be considered as two local pitch periods for both voiced and unvoiced frames. For voiced frames the local pitch is the pitch of the frame itself, where as for the unvoiced frames the local pitch is the last modified pitch.

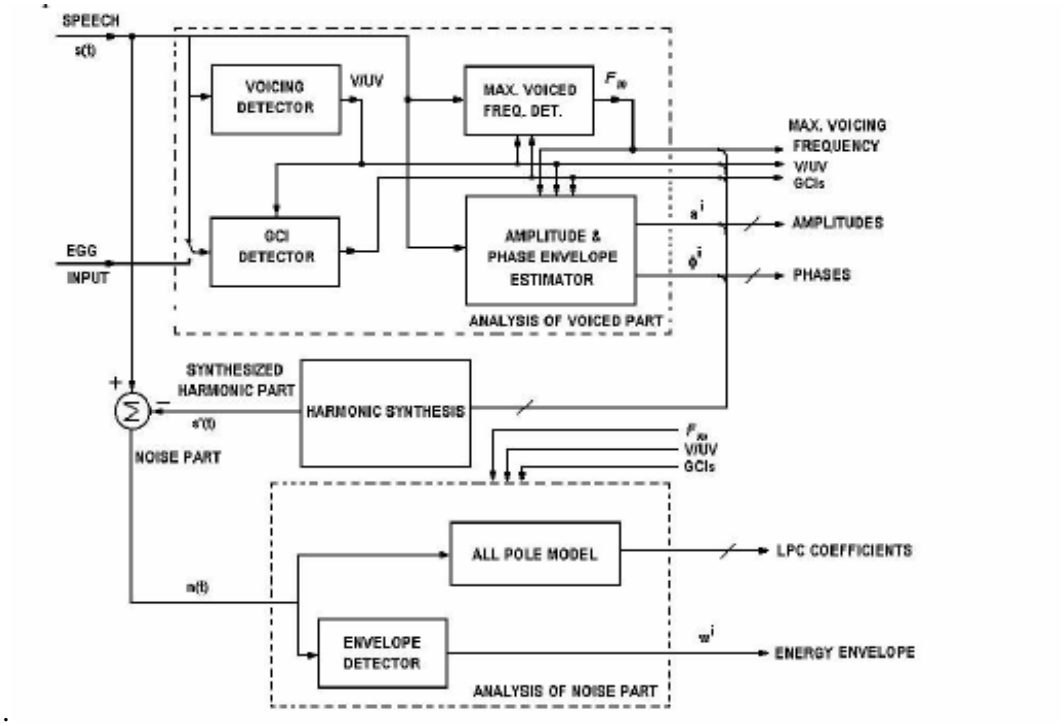

Fig.3. Analysis of speech using HNM from [3]

4.2. Synthesis of speech using HNM

By applying the parameters that are estimated using the analysis of speech to the synthesis scheme completes the speech synthesis part. The scheme for synthesis of speech [3] is shown the figure below. The parameters obtained during the analysis on frame-by-frame basis should be interpolated before synthesis for obtaining the parameter values at each sample. Generally we use the linear interpolation technique to obtain the parameters at each sample. But before the interpolation of the phase values it is necessary to carry out phase unwrapping. The sum of the harmonics after the linear interpolation of the amplitudes and phases (after phase unwrapping) gives the harmonic synthetic part**.**

By multiplying the interpolated energy envelope function with the LPC filter output (to which the LPC coefficients and the white gaussian noise signals are applied) we get the synthesized noise part of the HNM. It is important to note that the synthesized harmonic part will be used

during the analysis of speech for obtaining the noise part. By adding the synthesized harmonic part and synthesized noise part we get the synthesized speech. For estimating the glottal closure instants it is efficient to use the electroglottogram waveform from an impedence glottography.

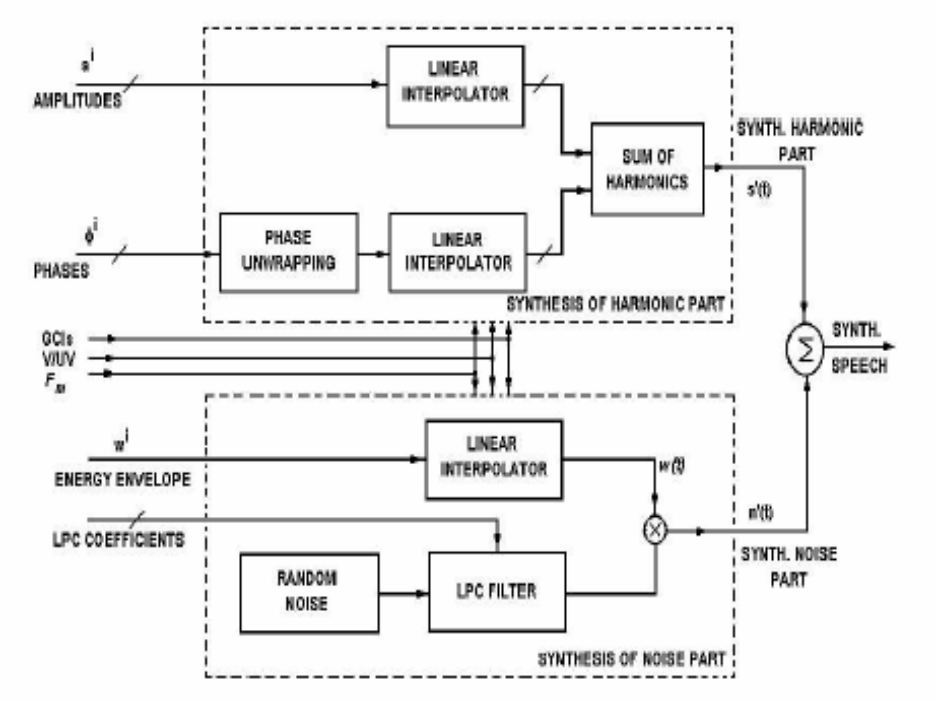

Fig.4. Synthesis of speech using HNM from [3]

4.3. Different ways to generate harmonic signal

There are four different techniques [4] for the generation of the harmonic signal using HNM. This is important for reducing the complexity of HNM, because more than 80% of the execution time of the HNM synthesis module is spent on generating the harmonic signal.

4.3.1. Straight Forward synthesis, SF

In this method the synthetic signal is generated directly by applying Equation (7). Hence the name called Straight Forward Method. The main problem with this method is the generation of cosine functions, which is very expensive. In this paper only this method is discussed.

4.3.2. Inverse Fast Fourier Transform, IFFT

The first thought to speed up the generation of synthetic signal is the use of inverse FFT [4]. FFTs may be used when the number of frequency bins (size of the FFT) is a number of a power of two. Because the number of harmonics may not be such a number, and hence it is necessary to assign the known frequency information (harmonics) to the closest frequency bins. This introduces, however, an error in the synthetic signal. Bigger size of FFT will cause smaller error (or, otherwise, higher SNR). However, bigger size of FFT will slow down the generation of the signal (higher complexity). McAulay and Quatiery found that for 4 kHz bandwidth of speech no loss of quality was detected provided the FFT length was at least 512

points. The bandwidths of the order of 8 kHz are also not enough for getting minimum error. Therefore, with larger FFT sizes (e.g., 1024, 4096, 8192), it is possible to reduce an error in the synthesized signal by increasing the size of the FFT. Hence the generation of the harmonic signal becomes slow considerably.

4.3.3. Recurrence Relations for cosine functions, RR

Trigonometric functions whose arguments from a linear sequence $\theta=\theta_0+n\delta$ with n=0,1,2… are efficiently calculated by the following recurrence:

$$\cos(\theta+\delta) = \cos\theta - [\alpha\cos\theta + \beta\sin\theta] \tag{9}$$

$$\sin(\theta+\delta) = \sin\theta - [\alpha\sin\theta - \beta\cos\theta] \tag{10}$$

where α and β are pre computed coefficients

$$\alpha = 2\sin^2(\frac{\delta}{2}) \tag{11}$$

$$\beta = \sin\delta \tag{12}$$

When the increment δ is small, then the recurrence relations do not lose significance. For each harmonic, k, we have to compute the coefficients α_k and β_k for $\delta_k=k\omega_0$.

4.3.4. Delayed Multi-Resampled Cosine function, DMRC

In this method the phase information will be first transformed into phase delays. The phase delay, t_k, of the k-th harmonic is defined as:

$$t_k = -\phi(k\omega_0)/k\omega_0 \tag{13}$$

where $\phi(k\omega_0)$ represents the measured phase at $k\omega_0$ frequency. Phase delays are expressed in samples and therefore are less sensitive to quantization errors. Transforming phase spectrum into phase delays allows us to write Eq.7 as following:

$$h(t) = \sum_{k=1}^{K(t)} A_k(t) X([tk - t_k] \bmod T) \tag{14}$$

where 'mod' stands for modulo, T is the integer pitch period in samples, and X denotes the cosine function:

$$X(t)=\cos(t\omega_0),\ t=0,1,\ldots T-1 \tag{15}$$

Eq.15 shows that $h(t)$ may be generated in a simple way. First we compute the signal $X(t)$ (actually, $X(t)$ is performed as there is limited number of integer pitch periods and it is just loaded from the disk during the generation of the harmonic signal), and then for every k harmonic, $X(t)$ is delayed by t_k, and down sampled by a factor k.

4.4. Speech transformations using HNM

There are a variety of techniques to modify the speed, pitch, and spectrum of speech signal [1]. In this section time scale modification and pitch scale modification techniques using HNM are presented.

4.4.1. Time scale modifications using HNM

This section provides the time scale modifications using pitch-synchronous overlap-add method. The input signal *s(t)* is modeled as the sum of the harmonic part *h(t)* and the noise part *n(t).* Where *h(t)* is given by

$$h(t)= \sum_{k=-K(t)}^{K(t)} A_k(t)\exp(jkt\omega_0(t)) \tag{16}$$

where $A_k(t)$ is the complex harmonic amplitude at time *t*, $\omega_0(t)$ is the fundamental frequency and *n(t)* is the noise component. These parameters will be updated at specific time instants denoted by t_i. With in each frame $[t_i, t_{i+1}]$, the fundamental frequency $\omega_0(t)=\omega_0(t_i)$ is held constant; the complex amplitudes $A_k(t)$ are affined functions of time:

$$A_k(t)=a_k(t_i)+(t-t_i)b_k(t_i) \tag{17}$$

$a_k(t_i)$ is the original complex amplitude of the *k*th harmonic in time-frame *i*; it represents the original amplitude and the phase of the harmonic at the time instant t_i. $b_k(t_i)$ is the complex slope of the harmonic; $b_k(t_i)$ reflects pseudo linear variations of the harmonic amplitude and slight misadjustments of its instantaneous frequency. *k(t)* represents the time-varying number of pitch-harmonics included in the deterministic part. The noise part *n(t)* is supposed to have been obtained by filtering a white gaussian *g(t)* by a time-varying, normalized all pole filter *A(t, Z)* and multiplying the result by an energy-envelope function *w(t)*:

$$n(t)=w(t)[A(t, Z)*g(t)] \tag{18}$$

The first step for the time scale modification [1] is to determine from the stream of analysis time instants t_i and the desired time scale modification factor $\tau(t)$ an integer-valued function $\phi(i)$ which specifies the number of synthetic pitch periods that need to be generated from the set of parameters at time t_i. For example for a constant scaling factor of 1.5 the function $\phi(i)$ is 2 for odd values of *i*, and 1 for even values of *i* (two periods are generated from the odd numbered analysis time frames, and from the even-numbered analysis time frames). This operation is very similar to that involved in the PSOLA synthesis. From the stream of analysis time instants t_i' and corresponding synthetic short term signals, the synthetic time instants t_i' are recursively calculated according to

$$(t_{i+1}'-t_i')=\phi(i)(t_{i+1}-t_i) \tag{19}$$

Two different schemes are used to modify the harmonic and noise parts. The harmonic part will be obtained in the following way; $\phi(i)$+1 periods of signals are generated according to the synthesis formula

$$h(t)= \sum_{k=-K(t)}^{K(t)} B_k(t)\exp(jkt\omega_0(t)) \tag{20}$$

where the time-varying harmonic complex amplitudes $B_k(t)$ are now given by

$$B_k(t)=a_k(t_i)+\frac{b_k(t_i)}{\phi(i)}(t-t_i) \tag{21}$$

Notice that the slopes that represent the slow variations of the periodic structure of the harmonic part have been divided by $\phi(i)$. That is, the hypothesis of harmonicity, and the fact that an integer number $\phi(i)$ of pitch-periods are generated guarantee that the amplitudes and phases of the synthetic harmonic part at the instant t_{i+1}' are to be same as that of the original signal at previous instant t_{i+1}. The simplicity of the synthesis scheme stems from the fact that the both the analysis and synthesis are performed at a pitch synchronous rate (no specific phase correction is needed, as opposed to the method recently proposed by McAulay and Quatiery).

The noise part will be obtained as follows. We first synthesize a time scaled noise signal by filtering a unit variance white Gaussian noise through a time varying normalized lattice filter whose coefficients $k(t)$ are derived from the stream of analysis filters $A(t_i, Z)$ and the stream of synthesize time instants; this time, the reflection coefficients $k_i(t)$ between the synthesis time instant $[t_i', t_{i+1}']$ will be obtained by linearly interpolating the reflection coefficients of the models $A(t_i, Z)$ and $A(t_{i+1}, Z)$. The time domain energy envelope function is then time scaled using a PSOLA-like technique [7]. This time-scaled envelope is finally to be applied to the noise signal, yields the time-scaled noise component.

The main advantage of this approach is to eliminate most of the artifacts encountered with PSOLA (especially when large time-stretching factors are used). In those methods, time stretching of voiced portions of speech is achieved (explicitly or implicitly) by replicating the same short-term signals over successive pitch periods. The noise part undergoes the same replication, a process that introduces an artificial periodicity resulting in a 'metallic' sound quality.

4.4.2. Pitch scale modifications using HNM

Pitch scaling in HNM [3] can be carried out by synthesizing the speech with interpolated original amplitudes and phases at the multiples of the scaled pitch frequency, which results in an unnatural quality and for obtaining natural quality output frequency scale of the amplitudes and phases of the harmonics of the original signal are needed to be modified by a speaker dependent warping function. Hence, it is necessary to study the relation between the pitch frequency and the vocal tract parameters.

The scheme for pitch scaling is shown in Fig.5. First the parameters of the speech of the source speaker should be calculated and then these parameters should be modified for achieving the target pitch contour using a warping function. The warping function will be obtained by studying the relationship between pitch frequency and formant frequencies for the vowels spoken at several notes. Then the re-synthesis will be performed from the modified parameters.

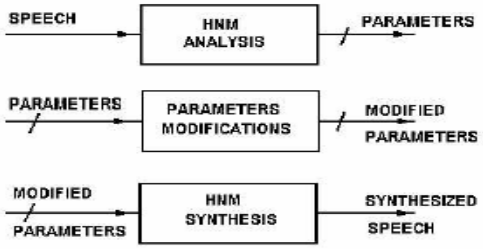

Fig.5. Scheme for pitch scale modification from [3]

CONCLUSION

HNM, a model for speaker characteristic modification has been presented. I conclude that, the analysis method, using HNM, discussed above can make it possible to accurately estimate the harmonic part, which can simply be subtracted from the original signal in the time domain to obtain the noise part. Methods for time scale modification and pitch scale modification, using HNM, have been presented. Since the model is parametric, it is possible to modify specific qualities of the speaker voice.

Acknowledgement

I wish to express my sincere gratitude to Prof. P.C.Pandey for his constant guidance throughout the course of the work and many useful discussions, which enabled me to know the subtleties of the subject in proper way.

References

[1] J. Laroche, Y. Stylianou, and E. Moulines, "HNS: Speech modification based on a harmonic+noise model", in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing'93* Minneapolis, MN, Apr.1993, pp550-553.

[2] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis", *IEEE Trans. on Speech Audio Processing,* Vol. 9, no.1, pp21-29, Jan. 2001.

[3] P.K. Lehana, and P.C. Pandey, "Harmonic plus noise model based speech synthesis in Hindi and pitch modification", *Proc. 18th International Congress on Acoustics*, ICA 2004, Kyoto, Japan, Apr.4-9, 2004, pp. 3333-3336.

[4] Y. Stylianou, "On the implementation of the harmonic plus noise model for concatenative speech synthesis", *ICASSP 2000*, Istanbul, Turkey, June 2000.

[5] B. Gold, and N. Morgan, S*peech and Audio signal processing,* John Wiley, New York, 2002.

[6] L.R. Rabiner, and R.M. Schafer, *Digital processing of speech signals,* Prentice-Hall, Englewood, N.J. Cliffs, 1978.

[7] H. Valbret, E. Moulines, and J.P. Tubach, "Voice transformation using PSOLA techniques", *Speech Communications*, Apr. 1992, vol. 11, pp. 175-187.

[8] R.J. McAulay, and T.F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation", *IEEE Trans. on Acoustics, Speech, Signal Processing*, 34(4), 1986, pp. 744-754.

[9] T.F. Quatieri, and R.J. McAulay, "Speech transformations based on a sinusoidal representation", *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-34:** 1449-1464, 1986.