# Table of Contents:

# Table of Figures

## 1. Abstract

This project deals with a problem of voice conversion: voice morphing.

The main challenge is to take two voice signals, of two different speakers, and create N intermediate voice signals, which gradually change from one speaker to the other.

This document will explore the different approaches to the problem, describe voice signal analysis techniques, and discuss a new approach to solving this problem.

The solution offered in this document makes use of 3-D surfaces, which capture the speaker's individuality (a surface from which it is possible to reconstruct the original speech signal). Once two surfaces are created (from two speakers), interpolation is applied in order to create a third speech signal's surface.

These surfaces are created by sampling the residual error signal from LPC analysis, and aligning the samples along a 3-D surface.

Once an intermediate error signal is created from the interpolation surface, a new speech signal is synthesized by the use of a new vocal tract filter.

The new vocal tract filter is obtained by manipulations on the lossless tubes' areas.

פרויקט זה עוסק בתחום מסוים הקשור להמרת אות דיבור: VOICE MORPHING. המטרה היא להמיר, באופן הדרגתי, אות דיבור של דובר אחד לזה של דובר שני, ע"י יצירת N אותות ביניים שישתנו באיטיות מהמקור ליעד.

במסמך זה נדון בקצרה בגישות השונות הנוגעות להמרת קול, נסקור שיטות לאנליזה של אותות דיבור, ונציע שיטה חדשה לפתרון הבעיה.

הפתרון המוצע בפרוייקט משתמש במשטחים תלת ממדיים הבנויים מאות השארית של אנליזת חיזוי ליניארי על מנת לאפיין דובר מסוים (מקור ויעד, משטחים שמהם ניתן לשחזר את אות הדיבור) . לאחר יצירת משטחים אלו, (עבור כל דובר בנפרד), נוצר משטח ביניים אשר מהווה בסיס לבניית אות דיבור שלישי.

ממשטח זה ניתן לשחזר אות שארית חדש שיהווה בסיס לסינתזת אות הביניים. הסינתזה תשתמש במסנן חדש שיבנה ע"י מניפולציות על שטחי השפופרות האקוסטיות.

## 2. The Project Goal

The Goal:

To gradually change a source speaker's voice to sound like the voice of a target speaker. The goal is not to make a mapping of one speaker's voice to another's, but rather to create N identical signals, which gradually change from source to target.

Applications:

One application of speech morphing could be for multimedia and for entertainment, for example, in voice morphing, like its facial counterpart (as often seen in video clips and TV commercials). While seeing a face gradually changing from one person to another's, the voice could simultaneously change, as well.

Another application could be for forensic use. For example, just as a sketch artist draws a suspect's face in the court, witnesses could be asked to describe a suspect's voice. This may be extremely difficult because it's not like describing a person's nose or eyes. In our case, it is suggested that the witness listen to a set of different voice sounds (from a "sound bank"), and identify which voice sounds are most similar to the one he or she heard at the scene of the crime. The next step would be to take an additional set (closer in sound to the one selected) and ask the witness to choose again.

At a certain point in this process, the use of this algorithm will be necessary in order to better zoom into the original voice heard.

The Challenges:

The first speaker's characteristics have to be changed gradually to those of the second speaker; therefore, the pitch, the duration, and the spectral parameters have to be extracted from both speakers. Then natural-sounding synthetic intermediates have to be produced. It should be emphasized that the two original signals may be of different durations, may have different energy profiles, and will likely differ in terms of many other vocal characteristics. All these complicate the problem, and thus, of course, the solution.

In figure 1, an example of two identical utterances of two speakers is presented;

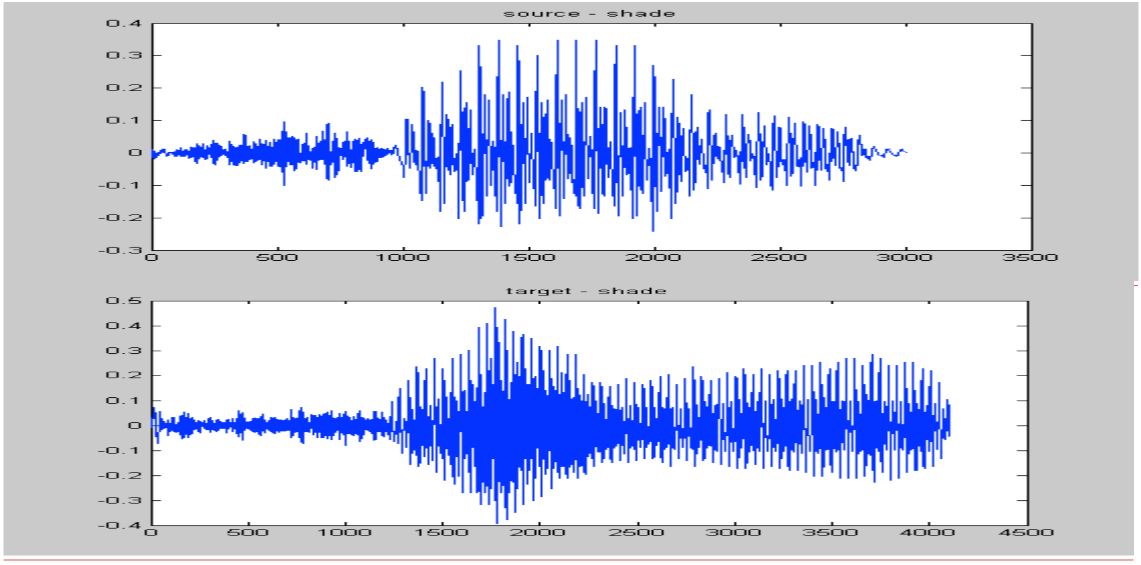


**Figure 1 – The word "shade". (a) – source , (b) – target.**

As can be seen, there are noticeable differences in shape, duration, and energy distribution.

# 3. Introduction

In this part of the document a few important aspects regarding speech signal analysis, speech modeling and speaker individuality will be presented.

## 3.1 Speech Signal Analysis

In order for us to offer a solution to the given problem, we must first understand the basics of speech signal analysis.

Voiced/Unvoiced/Silence determination:

A typical speech sentence signal consists of two main parts: one carries the speech information, and the other includes silent or noise sections that are between the utterances, without any verbal information.

The verbal (informative) part of speech can be further divided into two categories: (a) The voiced speech and (b) unvoiced speech. Voiced speech consists mainly of vowel sounds. It is produced by forcing air through the glottis, proper adjustment of the tension of the vocal cords results in opening and closing of the cords, and a production of almost periodic pulses of air. These pulses excite the vocal tract. Psychoacoustics experiments show that this part holds most of the information of the speech and thus holds the keys for characterizing a speaker.

Unvoiced speech sections are generated by forcing air through a constriction formed at a point in the vocal tract (usually toward the mouth end), thus producing turbulence.

Being able to distinguish between the three is very important for speech signal analysis.

Characteristic features for v/un determination

1. *Zero Crossing Rate*: The rate at which the speech signal crosses zero can provide information about the source of its creation. It is well known, as can be seen in figure 2, that unvoiced speech has a much higher ZCR than voiced speech [2]. This is because most of the energy in unvoiced speech is found in higher frequencies than in voiced speech, implying a higher ZCR for the former. A possible definition for the ZCR [2] is presented in equation 1:

[illegible]

1) | |

2. *Energy*: The amplitude of unvoiced segments is noticeably lower than that of the voiced segments. The short-time energy of speech signals reflects the amplitude variation and is defined [2] in equation 2:

2) .

In order for to reflect the amplitude variations in time (for this a short window is necessary), and considering the need for a low pass filter to provide smoothing, h(n) was chosen to be a hamming window powered by 2. It has been shown to give good results in terms of reflecting amplitude variations. [2]

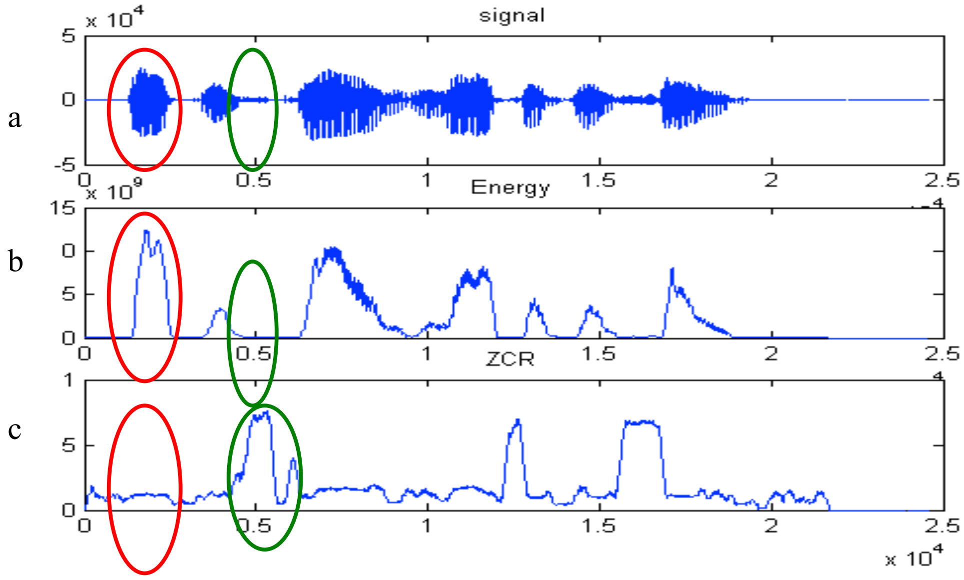


**Figure 2 – A speech signal (a) with its short-time energy (b) and zero crossing rate (c).**

In voiced speech (Fig. 2, red) the short-time energy values are much higher than in unvoiced speech (green), which has a higher zero crossing rate.

3. *Cross-correlation*. Cross-correlation is calculated between two consecutive pitch cycles. The cross-correlation values between pitch cycles are higher (close to 1) in voiced speech than in unvoiced speech.

Pitch detection:

Voiced speech signals can be considered as quasi-periodic. The basic period is called the pitch period. The average pitch frequency (in short, the pitch), time pattern, gain, and fluctuation change from one individual speaker to another. For speech signal analysis, and especially for synthesis, being able to identify the pitch is extremely important. A well-known method for pitch detection is given in [5]. It is based on the fact that two consecutive pitch cycles have a high cross-correlation value, as opposed to two consecutive speech fractions of the same length but different from the pitch cycle time.

The pitch detector's algorithm can be given by equations 3 and 4.

[illegible]

3) $\langle \quad \rangle$

4) [illegible] $\quad ; \quad \rho_\tau = \dfrac{\langle x, y\rangle}{\|x\| \cdot \|y\|} \quad ; \quad \|x\| = (\langle x, x\rangle)^{1/2}$

Figure 3 describes a vocal phoneme, in which the pitch marks are denoted in red.

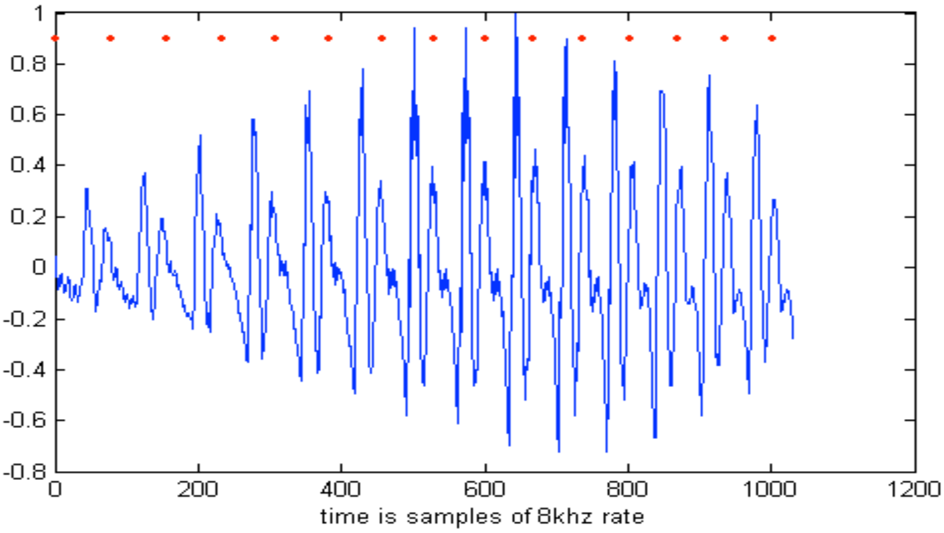


**Figure 3 – A phoneme with its pitch cycle marks (in red).**

## 3.2 Speech Modeling

The first step before understanding speech signal production and speaker individuality is to understand the basic model for speech production (which will be used in this project).

Lossless tube model:

Sound transmission through the vocal tract can be modeled as sound passing through concatenated lossless acoustic tubes. The areas of these tubes and the relationship between them force specific resonance frequencies.

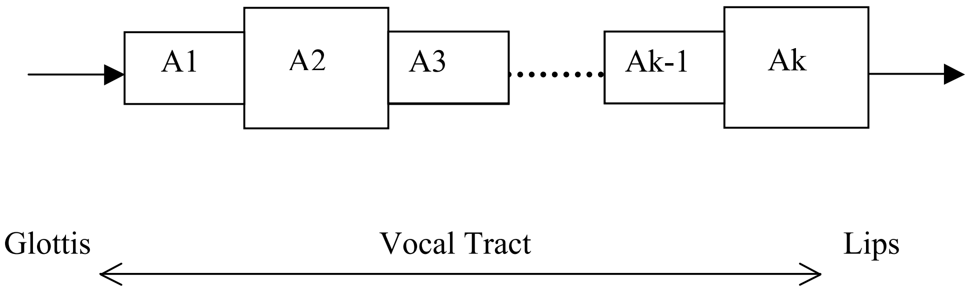


**Figure 4 – Lossless tube model**

It can be shown that this simple tube model can be implemented by a digital LTI filter, with a system function V(z) that provides vast information about the given speaker's characteristics. This filter, as will be shown later, is the basis of a commonly used method for speech synthesis (as applied in this project).
The mathematical relation between the areas of the tubes and of the filter's system function [3] can be seen in equation 5.

5)

The resonance frequencies of speech correspond to the poles of this system function. As will be discussed later, these resonance frequencies characterize a certain phoneme and speaker.

## Digital Speech Model:

In order for one to work with speech signals in a computer based (discrete) environment, a digital model for speech production is introduced.

*Excitation*: In voiced speech, the input for the vocal tract is the output of the glottis. The excitation for voiced speech is a result of opening and closing of the glottis (the opening between the vocal cords). This can be modeled as an impulse train passing through a linear system whose system function is G(z).
A commonly used impulse response for voiced speech [3] is the Rosenberg Model. The Rosenberg Model impulse response is shown below.

[illegible]

6) [illegible]

0

Unvoiced speech excitation is modeled by white noise.

*The Vocal Tract:* as discussed earlier, a digital linear filter can model sound transmission through the vocal tract.

*Radiation*: After passing through the vocal tract, the volume velocity flow finally reaches the lips. The pressure at the lips also needs to be modeled. This pressure is related to volume velocity by a high-pass filtering operation. A reasonable system function would be [illegible] [3].

*The Complete Model:*
In order to successfully synthesize a speech signal, a proper model that combines all of the parameters above is needed.

Figure 5 shows the complete digital model for speech signal production.

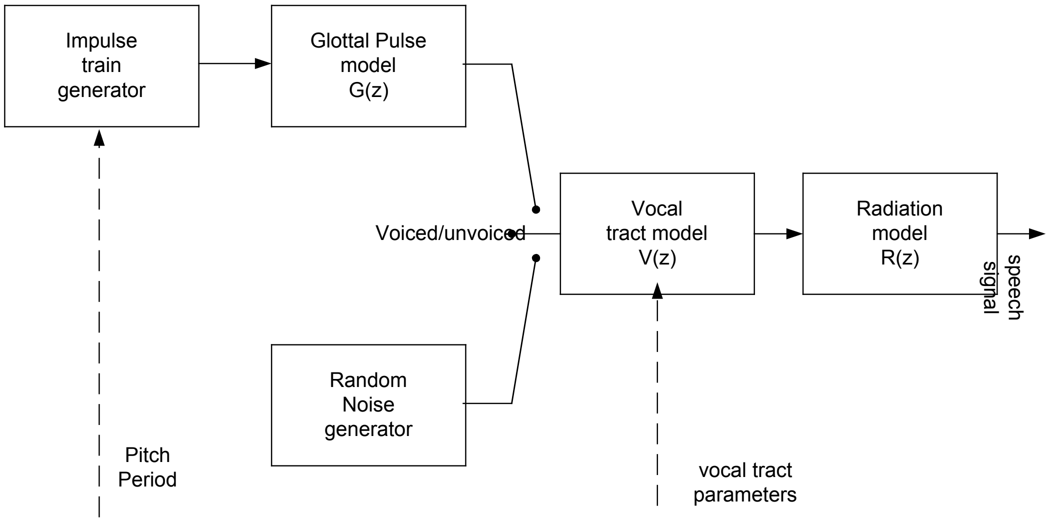


**Figure 5 – Discrete-time system model for speech production**

In the case of linear predictive analysis, the creation of a voiced or unvoiced signal involves the passage of the proper excitation through the linear system function:

7)

The area of the tubes (which will be used in this project's solution), and the vocal tract system function, can be easily calculated by the LPC technique, as described below.

## Linear Prediction of Speech

In this project, linear prediction (LP) is used for analysis and synthesis of speech. The LP is performed using the autocorrelation method of the12th order.
The linear prediction coefficients (LPC) are used to determine the vocal tract system function, the area of the tubes, and the signal error function (error signal).
An important by-product of the LPC analysis is the error signal, , which is the output of the prediction error filter [3]:

8)

where are the predictor's coefficients, and the input is the speech signal itself.
When referring back to the speech model, the error function is actually an approximation of the excitation.
Due to this, it is expected [2] that will have high values at the beginning of each voiced pitch period.
The error function is important for speech analysis and synthesis for two main reasons. First, by nature its energy is small, and thus better for speech coding purposes. Second, its spectrum is approximately flat, and therefore the formants' effects are eliminated. This is demonstrated in figures 7 and 6.

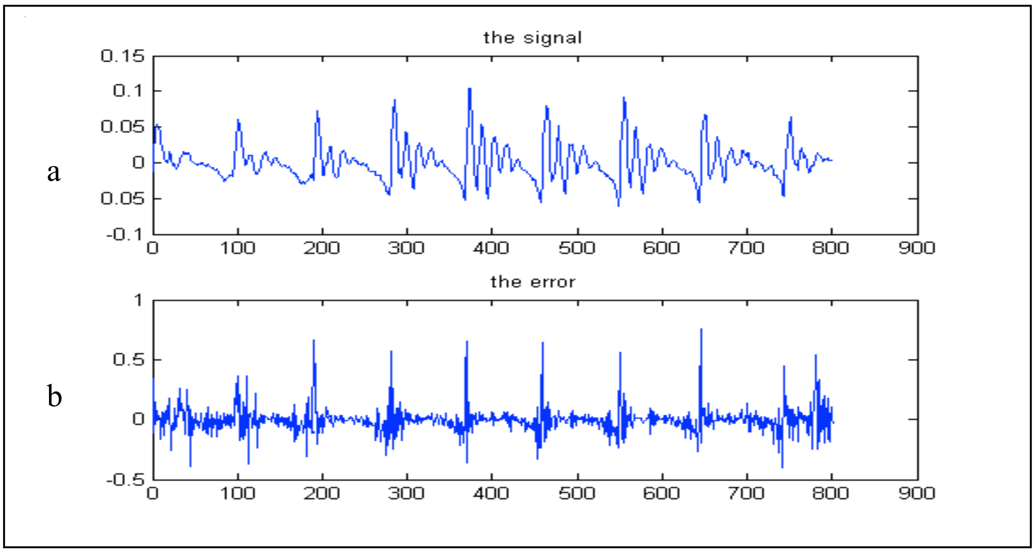


**Figure 6 –Voiced signal (a) and its error signal (b)**

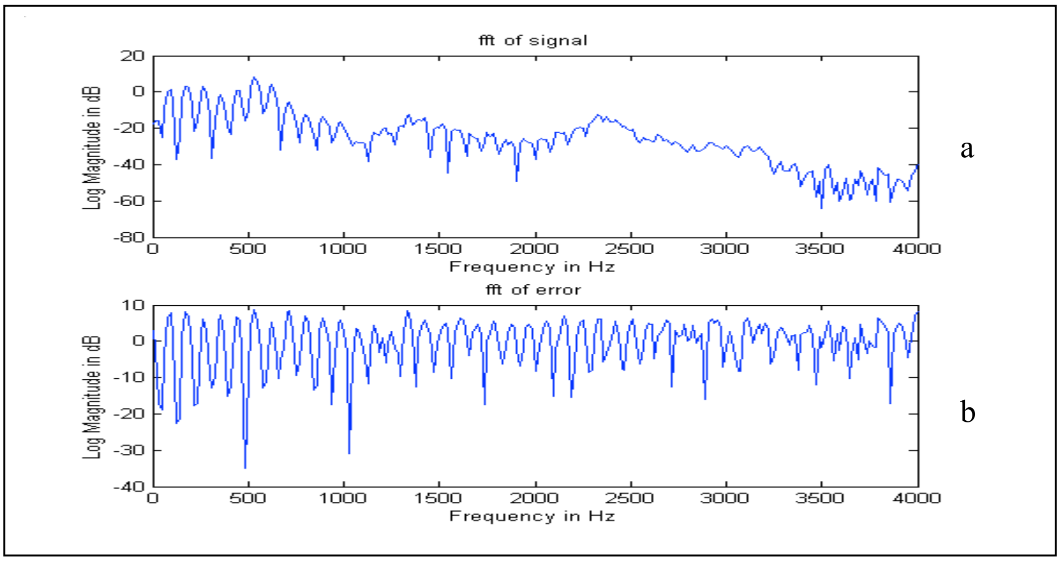


**Figure 7 –FFT of voiced signal (a) and the corresponding error signal (b)**

The fact that the error function's spectrum is relatively flat, leads to the assumption that manipulations on the residual error over the time domain will the degrade the speech utterance much less than manipulations on the vocal tract filter or on the speech signal itself.

## 3.3 Voice Individuality

Before trying to give a solution to the problem described in the project's goal, a better understanding of the different voice characteristics is needed. In this part of the introduction a few acoustic parameters, which are known to have the greatest influence on voice individuality are reviewed.

Acoustic parameters are divided into two groups: voice source (time dimension) and vocal tract resonance (frequency domain)[4].

Pitch frequency or fundamental frequency: Voice signals can be considered as quasi-periodic. The fundamental frequency is called the pitch. The average pitch period, time pattern, gain, and fluctuation change from one individual speaker to another and also within the speech of the same speaker.

Vocal tract resonance: The shape and gain of the spectral envelope of the signal (which is actually the vocal tract filter's frequency response), the value of the formants and their bandwidth.

Some research on voice individuality has concluded that the pitch (pitch fluctuation etc.) is the most important factor in voice individuality, giving formant frequencies the second place. However, other studies conclude that the spectral envelope has the greatest influence on individuality perception.

To conclude, it seems that there is no specific acoustic parameter that can alone define a speaker, but rather a group of parameters, with their respective importance vary from one individual to another, depending on the nature of speech materials. [4]

## 3.4 Voice Transformation

### Introduction

In this part, different aspects of voice transformation are discussed.

A perfect voice transformation system should take to consideration all the parameters discussed above. This is obviously very difficult and beyond our current capabilities. However many studies have been made on different approaches to the problem [6],[7],[8],[9],[10], all of them consisting of the same basic block diagram, as shown in figure 8.

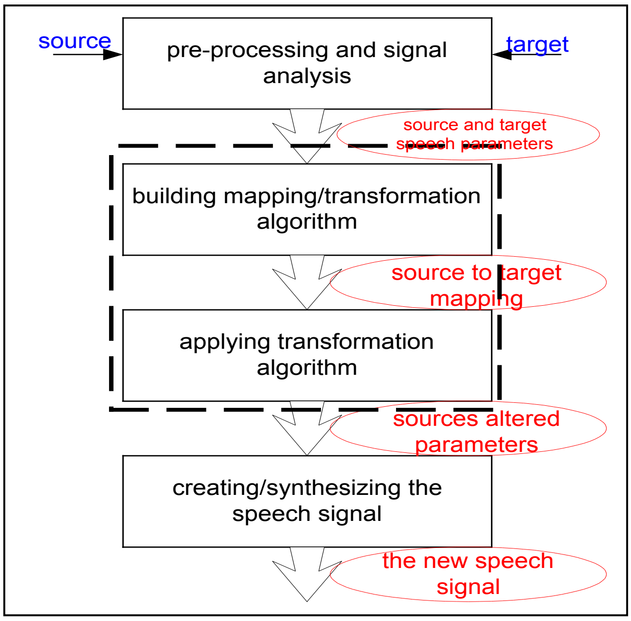

source
target
pre-processing and signal analysis
source and target speech parameters
building mapping/transformation algorithm
source to target mapping
applying transformation algorithm
sources altered parameters
creating/synthesizing the speech signal
the new speech signal

**Figure 8 – Basic voice conversion block diagram**

In the analysis stage of the conversion, typical individual parameters (according to the transformation algorithm) are calculated/evaluated. These parameters are necessary for future work. Such parameters might be pitch period duration, V/UN decision, LPC parameters, and the like.

In the second and third stages of the voice conversion, the mapping function between source and target is created and applied. Such mapping can be created, for example, by training a neural network [7] or by building a codebook [10] that maps the source parameters to those of the target.

Once all the relevant parameters are altered, the transformed speech is synthesized. Such synthesis might be, for example, the one described above (in the speech modeling section of this document) after changing the LPC parameters.

## Different aspects

As explained in the "Voice Individuality" section two main acoustic parameters – pitch period duration and formant frequencies – define the individual speaker. Although many approaches to voice conversion/morphing can be applied, all of them are based on two main manipulations of the speech features: time domain changes (such as the PSOLA algorithms) and manipulation on the vocal tract model filter (such as offered in [7]).

Time domain changes – pitch synchronous:

Although several different approaches are applicable here, they all lead to the same result: changes are made to the basic pitch cycle, such as pattern or frequency, and to its reoccurrence (the number of pitch cycles per phoneme may differ from source to target). For example, one pitch period of the source speaker could be altered and then multiplied. (The latter makes use of the Dynamic Time Warping algorithm).

In this part, the PSOLA technique is introduced.

*The PSOLA (Pitch Synchronous Overlap and Add) technique [8]*: In the basic TD-PSOLA (Time Domain PSOLA) system, prosodic modifications are made directly on the speech waveform. The same approach can also be applied on the error signal, resulting from the LPC analysis.

In the first step (analysis), pitch marks are calculated and denoted on the error signal, which is then divided into a stream of short-term signals, synchronized to the local pitch cycle.

The second stage is to warp the reference sentence in order to align it with the target sentence. Once this is done, the sequence of short-term signals of the reference speaker is converted into a modified stream of synthesized short-term

signals synchronized on a new set of time instants. This new set of time instants is determined in order to comply with the desired modifications. A new error signal is then obtained by overlapping and adding the new stream of synthesized short-term signals.
The last step is to synthesize the synchronized signal to the new pitch marks.

Manipulations of the vocal tract filter:

Moving formant frequencies (poles of the vocal tract filter) around the unit circle (especially those closer to the imaginary axes) results in loss of voice individuality. This, along with bandwidth and gain changes, can be achieved through direct changes to the filter, or changes to the area of the lossless tubes.
Here, a "speech morphing" technique that employs changes in spectrum parameter and fundamental frequency is presented [9].

*Spectrum parameter modification algorithm:* The first stage of this algorithm (analysis) is to find the pitch marks of each speaker's utterance and to create a correspondence table to match the source and the target.
In the second stage, values are set for the amount of modification desired. The third step is to make the necessary spectrum modifications, as follows:

1. Speech waveforms are extracted from source and target speaker's speech, by using the correspondence pitch table.
2. Spectra of the waveforms are calculated by FFT.
3. In the spectrum domain, a new set of FFT coefficients are generated by mixing FFT coefficients of both speakers. FFT coefficients below and above are copied from source and target speakers respectively. The threshold is set in the second stage.
4. Inverse Fast Fourier Transform (IFFT) is applied, to create the new intermediate waveform.
5. The TD-PSOLA algorithm is used to modify the fundamental frequency of the waveforms.
6. The final speech is created using the PSOLA algorithm.

In this project a combination of the two (time domain and formant frequencies manipulation) is proposed, in order to accomplish a successful gradual transition from the source to the target speaker.

# 4. The Proposed Solution – PWI

## 4.1 Introduction

Voice coding using Prototype Waveform Interpolation: PWI is a speech coding method described in [1]. It is based on the fact that voiced speech is quasi-periodic and can be considered as a chain of pitch cycles (as described earlier).

The slow change in pitch cycle shape and frequency suggests that sampling these cycles at constant time intervals should be sufficient in order to reconstruct the signal later on. This coding procedure can be followed for both the speech signal and its error function counterpart from the LPC analysis. The error function carries little energy (so it can be modeled by a delta impulse train), and this makes it a convenient choice for coding purposes. The coding technique is described in detail in [1] and will not be used in this project.

PWI speech morphing: As mentioned in [1] it is possible to represent a speech signal by a 3-D surface. The creation of such a surface will be described later on. Such a surface is the key for the speech signal's pitch and the shape of its waveform in time. In this project, the waveform surfaces of two different speech signals (from two speakers) merge together to create a new signal, which is their hybrid.

The error signal's spectrum is relatively flat, eliminating the effects of the formants; thus it is convenient to create its prototype waveform surface instead of the signal itself.

## 4.2 The Algorithm

### Concept:

As explained earlier, the goal is to create N intermediate signals between two different signals from two speakers. In the solution proposed, the fact that most of the speaker's individuality can be found in its voiced phonemes and that unvoiced speech carries little information (as well as being difficult to manipulate), leads to the decision that only the voiced sections are used in the morphing processes.

By using two prototype waveform surfaces (one for each speaker) for each phoneme, a third surface waveform, which is an interpolation between the two, is created. Than, an intermediate error signal is reconstructed from the intermediate surface. This error signal will be used to synthesize the intermediate phoneme.

In order to synthesize a voiced phoneme from the new error signal, a new LPC model is created from the two that already exist as will be described later.

## Basic block diagram:

A basic block diagram of the algorithm is presented in figure 9.

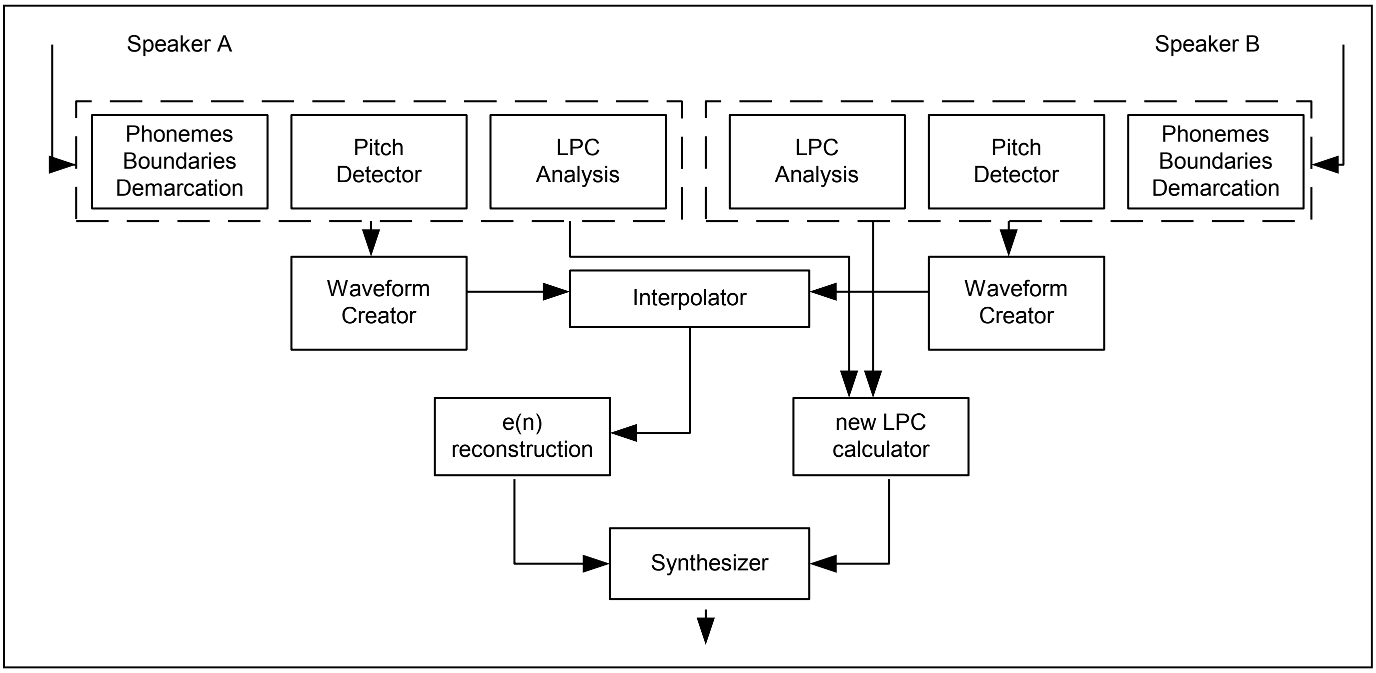


**Figure 9 – Project's Block Diagram**

Phoneme Boundaries Demarcation: This part is responsible for signing out the voiced phonemes and matching between those that belong to the source and those that belong to the target. This can be done both manually or by using zero crossing rates, energy, cross-correlation (for voiced speech identification) and Dynamic Time Warping for matching between the source and target phonemes.

Pitch Detector and LPC analysis: These two blocks are a part of the first layer of the diagram, which is the pre-processing stage. The error signal we will be working with is pitch-synchronous; it is a chain of the error signals of each pitch cycle (also the error signal of filtering the whole phoneme will be tested).

Computation of Characteristic Waveform Surface (waveform creator):

The surface is displayed along (t) – the time axis and (′ ) – the phase axis. The prototype waveforms are displayed along the phase axis, while the time axis displays the waveform evolution.

A typical prediction error signal (linear prediction residual signal) is shown in figure 10.

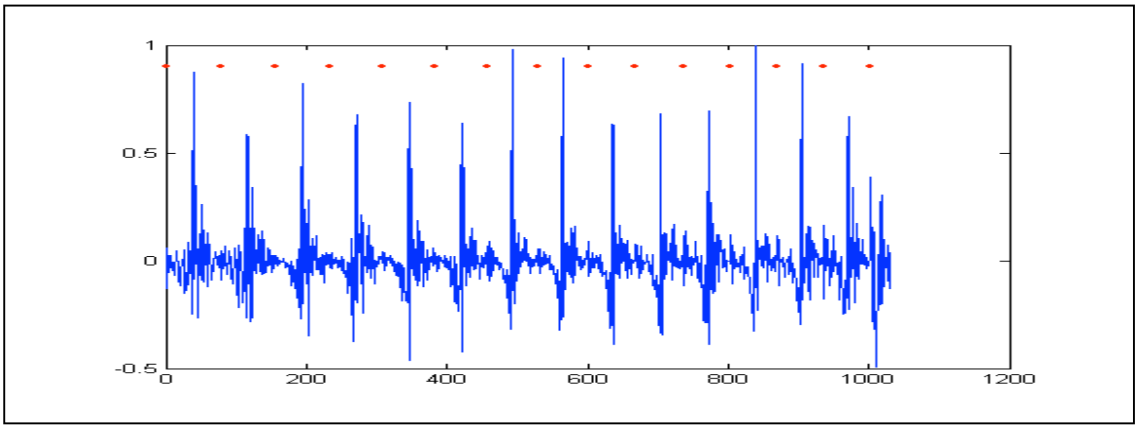


**Figure 10 – An error signal with pitch marks**

In the process of creating the characteristic waveform surface, segments of the linear prediction residual signal are aligned to create a 3D surface. These segments are sampled every 2.5 msec and are 1 pitch cycle long. These segments are called “prototype waveforms,” and their duration is the best compromise between time and frequency.

The surface is created for each phoneme separately, as follows:

1. LPC analysis is applied on the speech signal to create its residual error signal and the vocal tract parameters. The residual error function is the basis for computation of the characteristic waveform surface.

2. Pitch detection is applied in order to create a short-time pitch cycle function that will track the pitch cycle change through time. At any given point in time, the

pitch cycle is determined by an interpolation of the pitch marks obtained by the pitch detector.

3. Around a sampling time $t_i$, which changes at a rate of 2.5 msec, a rectangular window with a duration of one pitch period, multiplies the error function in order to create a prototype waveform.

4. In order to reconstruct the signal from the surface, it is extremely important to maintain similar and minimum energy values at both ends of the prototype waveform, (because these two samples actually represent the same place when referring to a periodic signal). This will allow a time sample in the window's center when sampling a prototype waveform.

5. Because the pitch cycle varies in time, each prototype waveform will be of different length. Therefore it is necessary to align all samples between and force them to have the same number of samples.

6. Once a prototype waveform is sampled and aligned between a cyclical shift around the axis is needed to create the best cross-correlation with the former prototype waveform; thus creating a relatively smooth waveform surface when moving across the time axis.

Given that is the 3D waveform, and that a given prototype is: , the aligned prototype is given as: when

7. After crossing the whole error signal, only the prototypes at given points in time are given. In order to create a surface, which reflects the error's pitch cycle evolution through time, an interpolation along the time axis is performed.

Figure 11 shows the basic block diagram of the waveform creator.

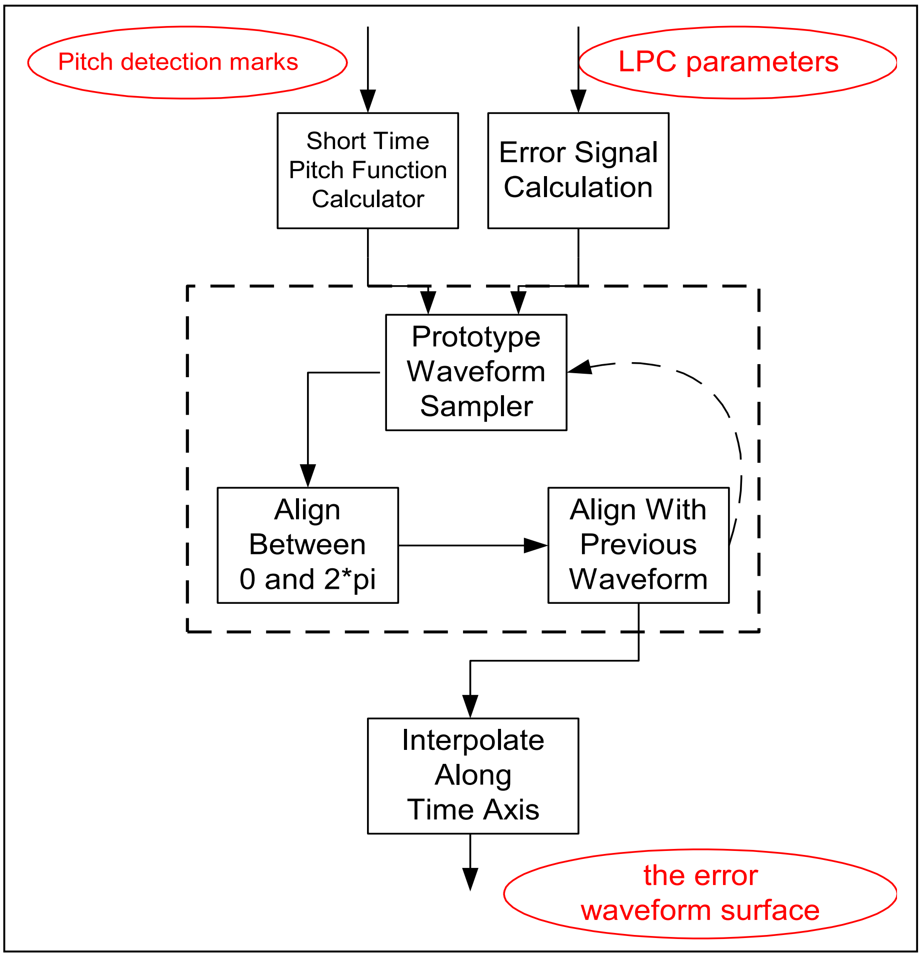
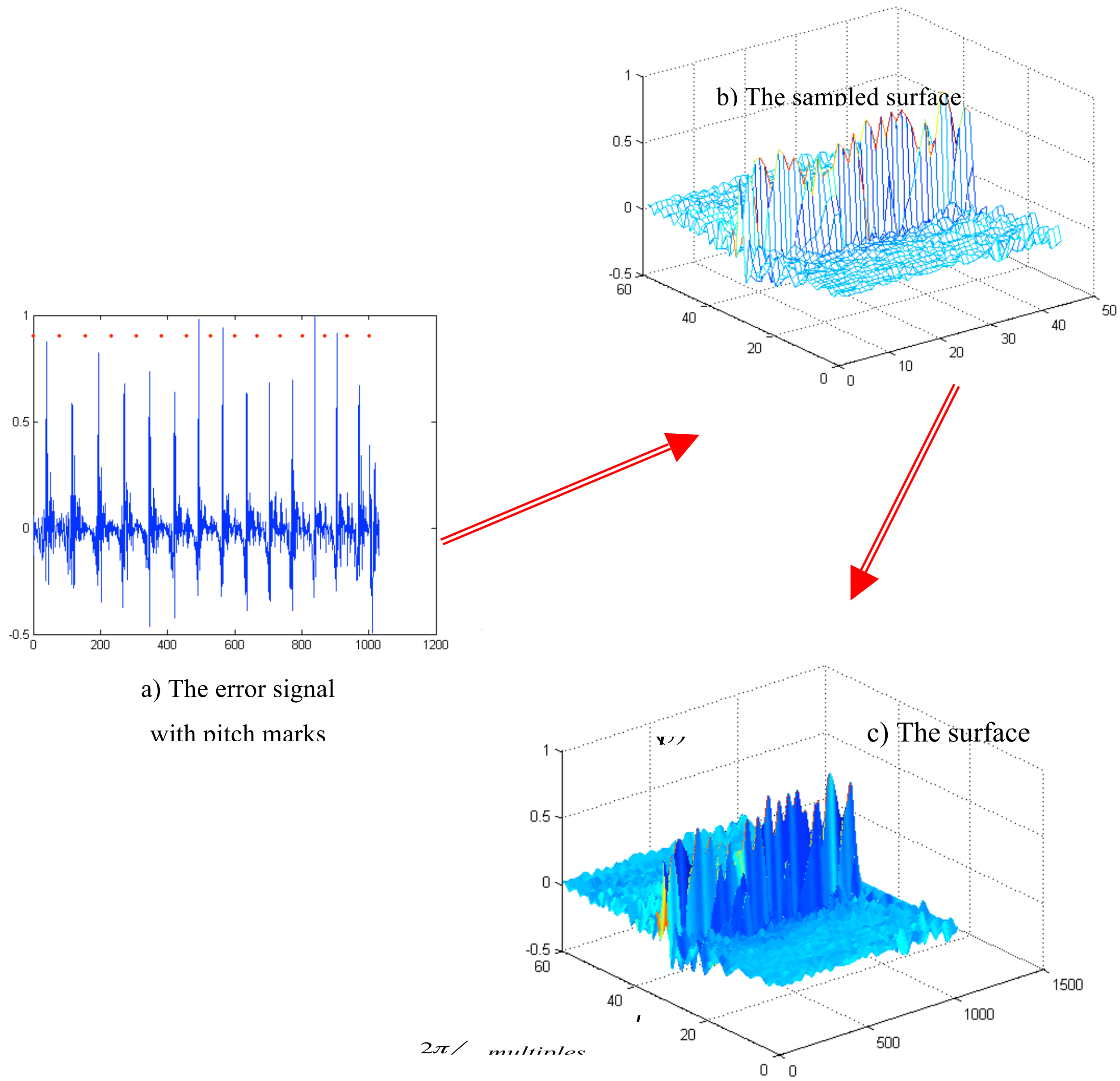

Pitch detection marks
LPC parameters
Short Time Pitch Function Calculator
Error Signal Calculation
Prototype Waveform Sampler
Align Between 0 and 2*pi
Align With Previous Waveform
Interpolate Along Time Axis
the error waveform surface

**Figure 11 – Waveform creator block diagram**

Figure 12 shows the three major steps of the surface creation.



a) The error signal with pitch marks

**Figure 12 – Waveform creation in 3 stages: (a) error signal with pitch marks; (b) sampled characteristic waveform; (c) final surface**

Reconstruction of the error signal:

The residual error signal from the prototype surface is reconstructed by defining a one-dimensional function, $\phi(t)$, which is the integral of the fundamental frequency over time, in other words, for a given , which solves:

and generally:

11) .

The function $\phi(t)$ is defined:

12)

Given that is the short-time pitch cycle of the speech signal, and that it changes linearly in time and is defined as: .

In figure 13, $\phi(t)$ along can be seen in red.

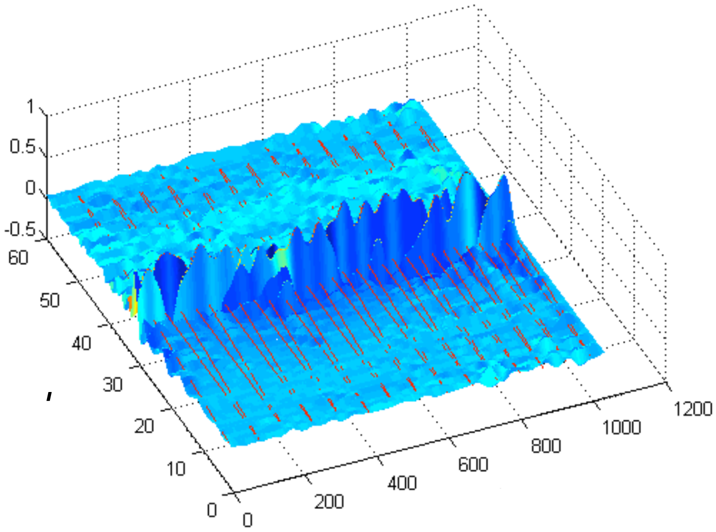


**Figure 13 – Error signal reconstruction using $\phi(t)$.**

Once $\phi(t)$ is defined, can be reconstructed using equation 11.

The Interpolator Morphing System:

Once the two waveform surfaces are created, both for the first and the second speaker for each voiced phoneme, Let be the PWI surface for the first (source) speaker and be the PWI for the second (target) speaker.

As described earlier, the new intermediate waveform surface will be, an interpolation of the two surfaces, and . Therefore:

13)

*$\alpha$ is the relative part of $u_s(t,\phi)$*

The last step in finding the new residual error signal (after creating ) is to construct it from the waveform surface.
The reconstruction is performed by defining: ,
where is created by the following equation:

14)

and by the hypothesis that the new pitch cycle of the intermediate signal changes linearly through time.

is calculated as an average of the source and target's short-time pitch function, as shown in equation 15:

15) [illegible]

New LPC creator and the synthesizer

It is well known that prediction parameters (i.e., the coefficients of the predictor polynomial A(z)) are very sensitive to quantization [3], because they are usually small and they are not defined on a linear space. There for their quantization may result in an unstable filter and an unrecognizable speech signal.

However, certain invertible nonlinear transformations of the predictor coefficients (and other manipulations, which will be tested here) result in equivalent sets of parameters that tolerate quantization better. An example of such a set of parameters is the PARCOR parameters, which are related to the areas of the lossless tubes, as shown in the equation 16:

16) [illegible] ——

This seems to indicate that in order to create a new set of LPC parameters that will define a new speech signal, an interpolation on the existing ones (source and target) should be sufficient.

For example, if the source and target are modeled as N lossless tubes with areas [illegible] and [illegible] respectively, then the new signal's vocal tract can be modeled by:

[illegible] .

After calculating the new areas, the prediction filter is computed and the new vocal phoneme synthesized according to the following schema:

1. Calculate new PARCOR parameters by reversing equation 16.
2. Calculate new LPC from PARCOR.
3. Once the LPC parameters are calculated, a reverse filtering is done on the new error signal in order to obtain the new speech signal.

Figure 14 outlines the basic block diagram for the creation new phonemes from the two surfaces.

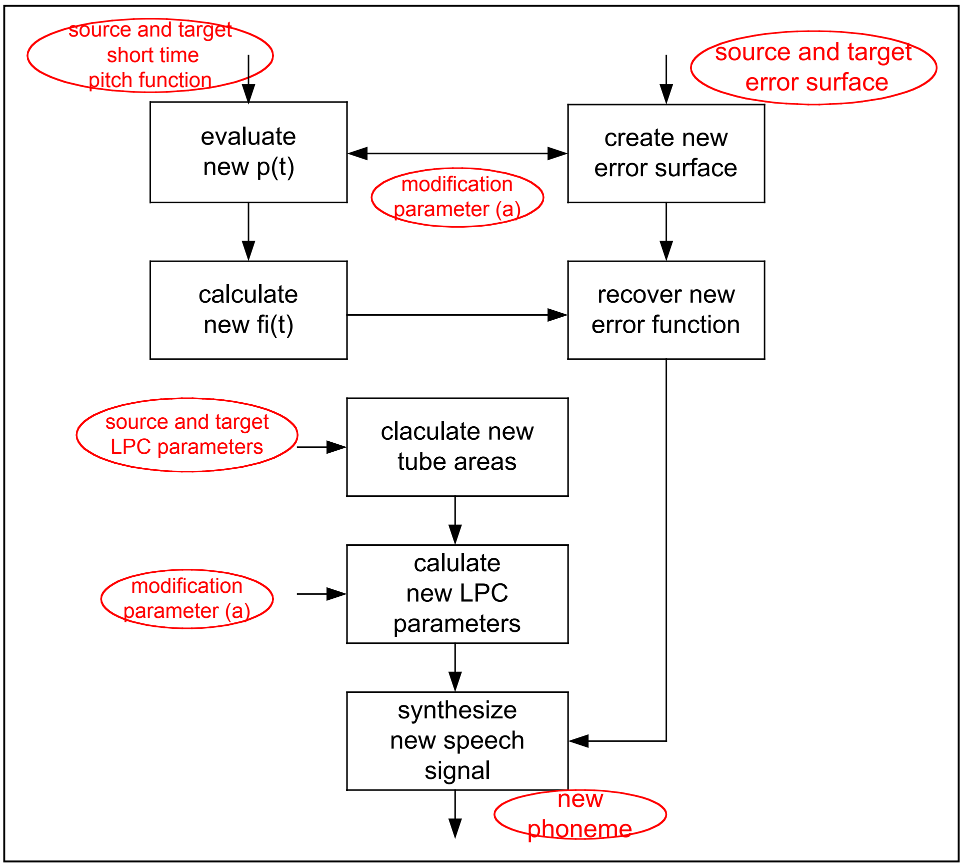


**Figure 14 – new signal creation block diagram.**

The final and new speech signal is created by cascading the new vocal phonemes in order along with the source's/target's unvoiced phonemes and silent periods.

# 5. Work Summery

## 5.1 Current Results

In this section the work that has been done until now will be reviewed

### 5.1.1 The environment

The basic knowledge and tools needed for the implementation of this algorithm has been gathered.

Basic functions for LPC analysis, pitch detector, speech signal synthesis (pitch synchronies and asynchronies), voiced/unvoiced/silence determination by ZCR, energy and cross-correlation have been built.

### 5.1.2 Algorithm's implementation

Computation of Characteristic Waveform Surface (waveform creator):

UP to now, a few basic waveform surfaces have been created for two speakers. The first step for the computation of these surfaces starts with the marking of voiced phonemes manually. The algorithm for computation of characteristic waveform surface was than applied on these signals. Figure 15 presents such a surface for a given phoneme.

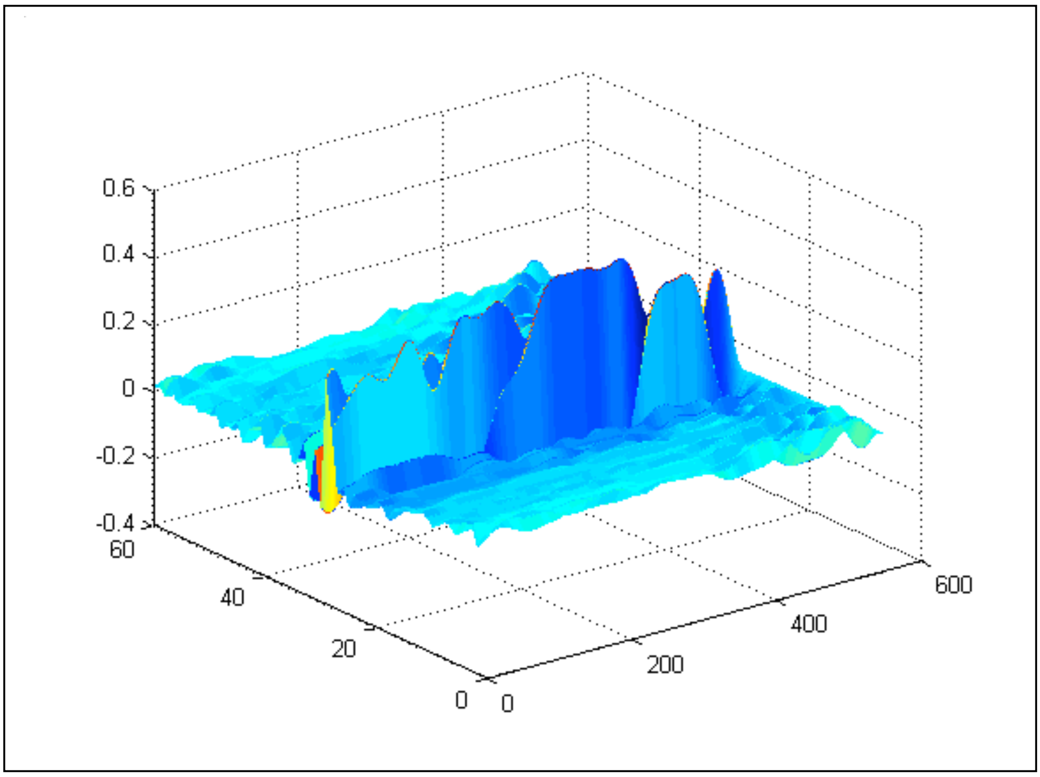


**Figure 15 – a characteristic waveform surface.**

However, as can be observed in figure 16, the implementation is not satisfying and further work is needed, due to the fact that some surfaces turn out to be not smooth enough across the time axis.

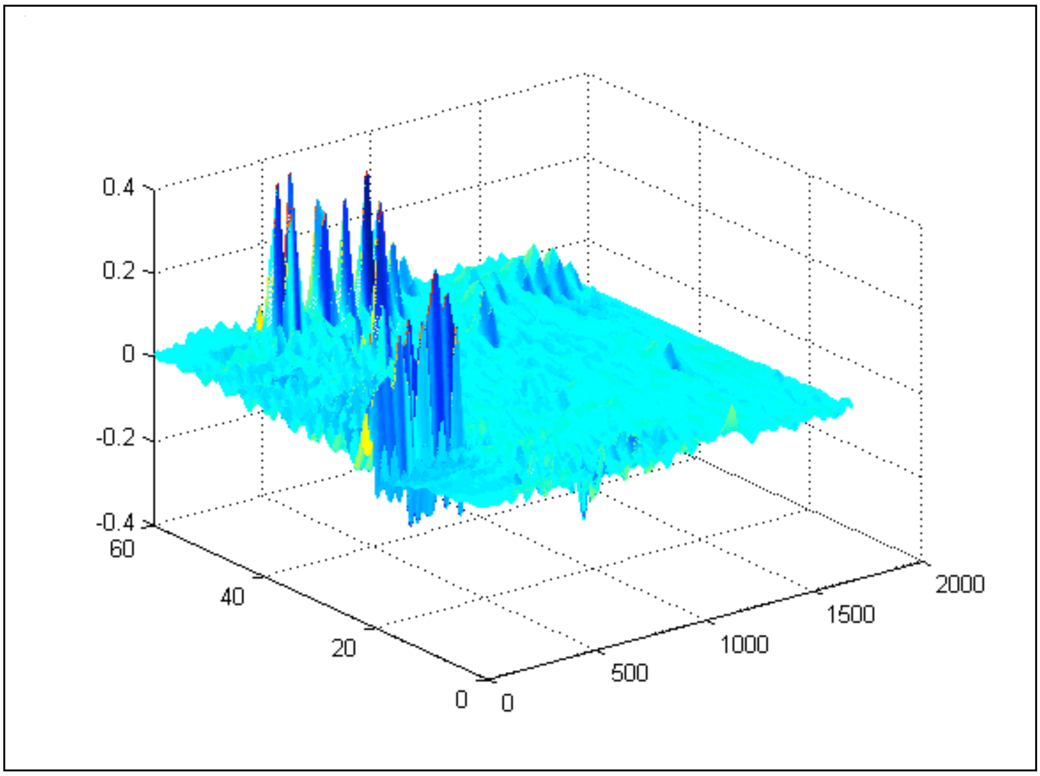


**Figure 16 – an example for a not yet finished surface.**

Reconstruction of the error signal from it's Characteristic Waveform Surface

Up to now, a function for the reconstruction of the error signal was constucted. This function seems to give good results (in a subjective listening test) for phonemes whose surface is smooth. However other phonemes' reconstructions give poor results, that means a hoarse voice.

## 5.2 Open issues

### 5.2.1 the environment

The main improvement that has to be done is in the pitch detector in durations that contain sudden jumps.

### 5.1.2 Algorithm's implementation

The next two sections are crucial to continue the development of the morphing algorithm.

Computation of the characteristic waveform surface (waveform creator):

. A good creation of a new, intermediate speech signal can be successfully done only after reliable surfaces, which reflect the two speakers will be created.

Reconstruction of the error signal from it's characteristic waveform surface

At this stage, only preliminary work was done. This part: depends on:

1. A good creation of the surface.
2. The creation of short-time pitch signal, which reflects the pitch evolution through time.
3. Calculation of $\phi(t)$.

The "Interpolator Morphing System" and "New LPC creator and the synthesizer"

As described in part 4 of this document.

Other issues

1. The nature of the synthesizer (and surface creator), in terms of a pitch-synchronized/unsynchronized synthesis has to be determined. The use of a pitch-synchronized system will not necessarily give the best results.
2. The creation of the short-time pitch contour can be achieved in different ways: linear interpolation, zero-order hold, sample-by-sample etc.

## 6. References

[1] "קידוד דיבור בקצבים נמוכים על בסיס מודל לזמן ארוך", אורית פלך– עבודת מגיסטר
למדעים, הפקולטה להנדסת חשמל, הטכניון חיפה, 2000.

[2] L. R. Rabiner , R. W. Schafer “Digital Processing of Speech Signals”

[3] J. H. McClellan, C. S. Burrus, A. V. Openheim, T. W. Parks, R. W. Schafer, H. W. Schuessler, “Computer Based Exercises For Signal Processing using Matlab 5,” Chapters 10,11.

[4] Hisao Kuwabara, Yoshinori Sagisaka, “Accustic characteristics of speaker individuality : control and conversion”, SPEECH COMMUNICATION 16 (1995) 165-173.

[5] Yoav Meden, Eyal Yair and Dan Chazan, “Super Resolution Pitch Determination of Speech Signals”, IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 39, NO 1, JANUARY 1991.

[6] Yannis Stylianou, Jean Laroche, Eric Moulines “High-Quality Speech Modification based on Harmonic + Noise Model”, Eurospeech – ’95 Spain.

[7] M. Narendranath, Hema A. Murthy, S. Rajendran, B. Yegnanarayana, “Transformation of formants for voice conversion using artificial neural networks” Speech Communication 16 (1995) 207-216.

[8] H.Valbret, E. Moulines, J.P. Tubach, “Voice Transformation using PSOLA technique”, Speech Communication 11 (1992) 175-187 North Holland.

[9] Masanobu Abe, “Speech Morphing by Gradually Changing Spectrum Parameters and Fundamental Frequency”, NTT Human Interface Laboratories.

[10] Levent M. Arslan ,“Speaker Transformation Algorithm using Segmental Codebook (STASC)”, Speech Communication 28 (1999) 211-226.