VOICE CONVERSION FOR RECONSTRUCTION OF DYSPHONIC SPEECH

by

Gönenç Seçil Tarakcıoğlu

B.S., Electrical and Electronics Engineering, Bogazici University, 2007

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Electrical and Electronics Engineering
Boğaziçi University

2010

ACKNOWLEDGEMENTS

I would like to gratefully and sincerely express my thankfulness to my supervisor, Prof. Levent Arslan, for his precious guidance, help and for sharing his unexcelled expertise with me. His felicitous advices gave me the confidence to carry on every time I encountered the struggles in the challenging task of dysphonic speech enhancement. I have taken and will always take him as an example in many aspects.

I deeply appreciate the genuine support and encouragement of my thesis committee members, Assoc. Prof. Murat Saraçlar and Prof. Sıddık Yarman. They contributed to this work in a lot of ways with their constructive comments, discerning ideas and their broad perspectives, also introducing new ways of looking at things in times of trouble. Being able to work with these three tremendous academicians has been a great luck and privilege.

I heartily appreciate the efforts of Prof. Emin Anarım, Prof. Ferhan Öz and Assist. Prof. Selçuk Köksal to bring me in touch with Gayem Köprücü who guided me in medical matters very patiently, friendly and open-heartedly. She became much more than my medical consultant from the very beginning.

I also thank to all the doctors at Ear, Nose and Throat (ENT) Clinic of Istanbul Faculty of Medicine, especially to Dr. Beldan Polat and Dr. Deniz Kanlıada, for the willingness, interest and helpfulness they showed towards my work; I thank to the larynx cancer patients for unhesitatingly volunteering during the data collection phase and to everyone who participated in the listening tests.

Special thanks to my very dear friends at BUSIM and to my equally dear colleagues at Sestek Inc one by one who shared, comforted and calmed. The joy I have with them and their creativeness always inspired and motivated me. I definitely cannot thank enough to Erinç Dikici, in particular, who miraculously never got tired of supporting me all the way miraculously and getting me back on track whenever I was

lost.

I am grateful to my professors and instructors at Bogazici University who -with a great wisdom, devotion and altruism- have evolved me into an engineer throughout my undergraduate and graduate years.

I also owe special thanks to TUBITAK - BIDEB for supporting me financially during my graduate studies in scope of the 2228 scholarship program.

Above all, I am mostly thankful to my mom and dad for their endless and unconditional love, care and for being the most amazing parents one can ever wish for. Without them none of this could have happened...

ABSTRACT

VOICE CONVERSION FOR RECONSTRUCTION OF DYSPHONIC SPEECH

The medical term "dysphonia" is used for voice disorders caused by functional impairment of voice organs which results in inadequate oscillatory movement of the vocal folds. Dysphonic speech is typically hoarse, rough and breathy with very weak phonation or with no phonation in severe examples (for instance, after total laryngectomy). Although there are speech therapy techniques as well as surgical and post-surgical rehabilitation methods practiced in medicine, a complete restoration of dysphonic speech is usually not possible with none of these elements.

This study addresses this specific problem and aims to computationally repair the voice source and vocal tract anomalies of dysphonic speech. The system proposed hereby builds upon the Source-Filter Speech Model and approaches the task twofoldedly: (i) Voice source repair involves substituting the missing glottal excitation and feeding it to the vocal tract filter according to the linear prediction equation. (ii) Vocal tract repair algorithm makes use of Voice Conversion principles and focuses on replacing the dysphonic speech parameters with those of normal speech. The performance evaluations have shown the developed reconstruction system to succeed in imitating some characteristics of normal speech, yet remarkable improvements cannot be achieved perceptually in terms of naturalness.

ÖZET

DİSFONİK KONUŞMANIN GERİÇATIMI İÇİN KONUŞMACI DÖNÜŞTÜRME

Tıbbi bir terim olan "disfoni", ses organlarındaki işlevsel bozuklukların yol açtığı ve ses tellerinin yetersiz salınım hareketi ile sonuçlanan ses bozulmaları için kullanılır. Disfonik konuşma tipik olarak boğuk, düzensiz ve solukludur; fonasyon çok zayıftır veya ağır örneklerde (örneğin, total larenjektomi sonrası) hiçbir fonasyon olmayabilir. Konuşma terapisi tekniklerinin yanı sıra, cerrahi ve ameliyat sonrası rehabilitasyon yöntemlerinin tıp alanında uygulanmasına rağmen, genellikle disfonik konuşma için tam bir yeniden inşa bu unsurların hiçbiri ile mümkün değildir.

Bu çalışmada, bu özel sorun ele alınmakta ve disfonik konuşmadaki ses kaynağı ve ses yolu anomalilerinin sayısal olarak onarımı amaçlamaktadır. Burada önerilen sistem Kaynak- Süzgeç Konuşma Modeli üzerine inşa edilmiştir ve problemi iki yönlü ele almaktadır: (i) Ses kaynağının onarımı, eksik olan gırtlak uyarımının yerine konulmasını ve doğrusal tahmin denklemi uyarınca ses yolu süzgecinin bununla beslenmesini içerir. (ii) Ses yolu onarım algoritması ise Konuşmacı Dönüştürme ilkelerini kullanmakta ve disfonik konuşma değiştirgelerinin normal konuşmaya ait olanlarla değiştirilmesine odaklanmaktadır. Başarım değerlendirmeleri, geliştirilen geri çatma sisteminin normal konuşmanın bazı özelliklerinin başarıyla taklit edilebildiğini göstermekle beraber doğallık açısından algısal olarak henüz kayda değer gelişmeler elde edilememiştir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS/ABBREVIATIONS

a_k	Complex pole of the vocal tract filter transfer function
$A_m(t)$	Amplitude of the m-th sinusoidal component in the sinusoidal model
$A(z)$	Transfer function of the linear prediction error filter
$e(n)$	Linear prediction error or the LP residual
E	Mean-squared prediction error
F_0	Fundamental frequency
G	Gain parameter
$G(z)$	Glottal shaping filter
$H(z)$	Vocal tract filter
p	LP order
$P(z)$	Symmetric LSF polynomial
$Q(z)$	Anti-symmetric LSF polynomial
$s(n)$	n-th sample of a speech signal
$\bar{s}(n)$	Approximation of the speech signal
$S(z)$	Signal spectrum
$u(n)$	Source excitation signal
$U(z)$	Source excitation signal spectrum
T_0	Fundamental period
α_k	Linear predictor coefficients
$\phi_m(t)$	Phase of the m-th sinusoidal component in the sinusoidal model
$\omega_m(\tau)$	Instantaneous frequency of the m-th sinusoidal component in the sinusoidal model
ABS/OLA	Analysis-by-Synthesis/Overlap-Add Model
DFT	Discrete Fourier Transform
EL	Electrolarynx
ENT	Ear, Nose and Throat

FFT	Fast Fourier Transform
HNM	Harmonic Plus Noise Model
IAS	Institutional Assessment and Studies
LP	Linear Prediction
LPC	Linear Prediction Coefficients
LSF	Line Spectral Frequencies
LSP	Line Spectrum Pair
MOS	Mean Opinion Score
MSE	Mean Squared Error
SR	Speech Recognition
TEP	Tracheoesophageal Prosthesis or Puncture
TFM	Transformation or Transformed Speech

1. INTRODUCTION

First, the motivation behind this study is discussed in this chapter, keeping the main goals and target directions in consideration. Next, relevant literature is shortly reviewed where the previous researches and studies related to this subject are summarized. Finally, the problem -which this work deals with- is stated. An outline of the thesis is also given at the end.

1.1. Motivation

Speech occupies a remarkable place in both daily and professional life. Although new means of communication keep being introduced consistently in today's technology driven world, verbal interaction still remains as the easiest and most common method of all. One of the IAS (Institutional Assessment and Studies) reports from the University of Virginia defines oral communication as "the effective interpretation, composition, and presentation of information, ideas, and values to a specific audience" [1].

The exchange of information between a speaker and a listener/listeners is only considered to be successful or efficient, if this information can be conveyed and understood entirely and correctly. Needless to say, voice production plays a significant role in ensuring these conditions at the speaker side. Accordingly, any voice deficiency -even temporarily- degrades the effectiveness of the so-called speech chain, while the situation is a lot worse for people suffering from long-lasting or permanent vocalization problems. Such a loss of normal voice can occur as a result of laryngeal cancer treatment. In this case, patients are typically only able to speak in hoarse whispers which are usually not easily perceptible. Despite the presence of certain voice rehabilitation and restoration techniques, this kind of speech still has much lower quality and intelligibility than normal speech. The patient's everyday living gets affected dramatically as a consequence, since most of the daily activities become into challenging tasks. In addition to other complications resulting from the disease, the physiological disorder may also bring with it social (self-)isolation, because the patient and the society is

usually uncomfortable with the voice defect.

Doctors, therapists and researchers have been pursuing alternative methods to enhance this low-quality speech with the hope of improving the patients' speaking capabilities while eliminating the psychological side effects as much as possible. This is the same motivation that the thesis presented hereby has been grounded upon.

1.2. Overview

The "Cancer Facts and Figures - 2010" Report of American Cancer Society stated that presumably 12,720 people would be diagnosed with cancer of the larynx in 2010 where the estimated number of new cases for all cancer types was 1,529,560 [2]. Due to this relatively low percentage of occurrence, the laryngeal cancer is not perceived as a threat in society as much as many other cancer types known. However, its physiological, psychological and sociological results have been keeping the multidisciplinary research on larynx cancer valuable and active.

The history of total laryngectomy dates back to 1874 (performed by Billroth). More and more scientists with increasingly varying interest areas have been involved ever since, and the attempts of diagnosing, analyzing and curing vocal impairments have grown in number especially after the first half of the twentieth century.

The earliest efforts such as [3, 4] tried to cognize and define "vocalization", the sub-segmental and supra-segmental clues of speech, followed by comparative studies [5-8] that strived for understanding the differences between normal and dysphonic speech both acoustical and phonetically.

First, the only way of speaking after laryngectomy was the speech by controlled belching (called esophageal speech by the laryngologist Seeman in 1919). The track of years following the first larynx surgery has gone along with evolvements in medicine and biomedicine: additional methods of post-laryngectomized speech production have emerged for surgically restoring the vocalization. Studies like [9-12] or [13]

investigated the vocal quality and the ability to convey pitch accent of these different methods. The results indicated that these speech alternatives can bring some features back, but the dysphonic speech still suffers under the lack of naturalness and intelligibility to ensure sustainable and painless communication. Finally, with the contribution of engineers and physical scientists in the past few decades, algorithms for speech reconstruction have also begun to be developed where some of them used formant manipulation [14, 15], some replaced the dysphonic speech segments by their normal equivalents using pattern recognition approaches [16] or Mixed Excitation Linear Prediction based approaches [17] while Pozo employed voice conversion algorithms to convert deviant parameters and regenerate speech with more natural quality and higher intelligibility [18].

1.3. Problem Statement

The effects of apparent laryngeal deficiencies (e.g. removal of larynx) are not restricted to defects in pitch generation, but they commonly change the whole speech production mechanism substantially. Consequently, the speech is altered to a great extent and loses many of its distinctive characteristics. So, the problem of voice repair can readily be interpreted as a voice conversion task where the new speech must be modified so that it resembles to the original or to a more natural sounding speech.

In this work, non-surgical dysphonic speech modification and enhancement systems have been proposed for individuals with laryngeal disorders to regain their ability of speaking with higher intensity and intelligibility. The work is hoped to ease the oral communication especially in cases like phone conversations where no nonverbal clues (gestures, mimics, lip movements, etc.) are present to validate or clarify the intended message. Most significant contributions of this thesis are

- two analysis-transformation-synthesis schemes for dysphonic speech modification to produce speech with vocal characteristics close to the original speech
- voice source substitution techniques proposed to replace the missing excitation in dysphonic speech and

- the qualification of their validity in different dysphonic speech cases.

1.4. Thesis Outline

The rest of this thesis is organized as follows:

Chapter 2 is an overview of speech theory and describes the mechanisms responsible for speech production. This section also includes background information on approaches regarding speech modeling.

Chapter 3 deals with the concept of "dysphonia" more intensely. Physiological and pathological reasons that lead to the loss of vocalization are investigated. Finally, different restorative options for dysphonic speech rehabilitation are presented and characterized.

Chapter 4 introduces the algorithms which have been employed throughout this thesis. A system to enhance the dysphonic speech has been developed based on the principles of voice conversion and speech re-synthesis. Two different schemes in time and frequency domains, respectively, are presented.

Chapter 5 is dedicated to tests and evaluations. Following an explanation of the test setup and test material, proposed schemes and their performances in terms of naturalness and intelligibility are examined by applying the implementation to real dysphonic cases. The evaluations consist of perceptual listening tests with some additional objective indicators.

Chapter 6 derives final conclusions from the work presented and brings this thesis to completion with suggestions for future research.

The list of utterances which constitute the speech corpora for the tests is also present in Appendix A as a supplementary piece of information.

2. SPEECH THEORY & BACKGROUND

The urge to adequately understand, analyze and model speech makes speech-related engineering a multidisciplinary area of research which is in close contact with sciences like acoustics, linguistics, phonetics or anatomy. The vocal production of speech is explained through assumptions based on acoustics and anatomy, as well. Accordingly, general concepts of speech production are covered in this chapter. These physiological aspects provide a basis for the speech models in literature: Sinusoidal Models and the Source-Filter Model. The Source-Filter Model is issued slightly more intensively here, with the details of model parameter estimation, since some of these tools and algorithms are also utilized in this thesis.

2.1. Theory of Speech Production

Anatomically, speech production is the process of generating an acoustic sound pressure wave with the participation of various structures in the human speech production system.

The vocal system is comprised of large components including *lungs* (source of air/energy along with the diaphragm), *trachea* (windpipe), *larynx* (organ where voice is produced), *pharyngeal cavity & oral (or buccal) cavity* (throat and mouth, respectively, usually grouped into the "vocal tract"), *nasal cavity* (nose, often called as the "nasal tract") and finer components which are called *articulators* (vocal cords or vocal folds, soft palate or velum, uvula, tongue, teeth, lips, jaw etc).

While speaking, the air flow from the lungs is forced through the glottis at the larynx to the pharyngeal, oral and nasal cavities. From the oral and nasal cavities it finally radiates through the mouth and nose, respectively.

The V-shaped opening between the vocal cords, called the glottis, has an especially significant role in speech production: it is responsible for modulating the air flow,

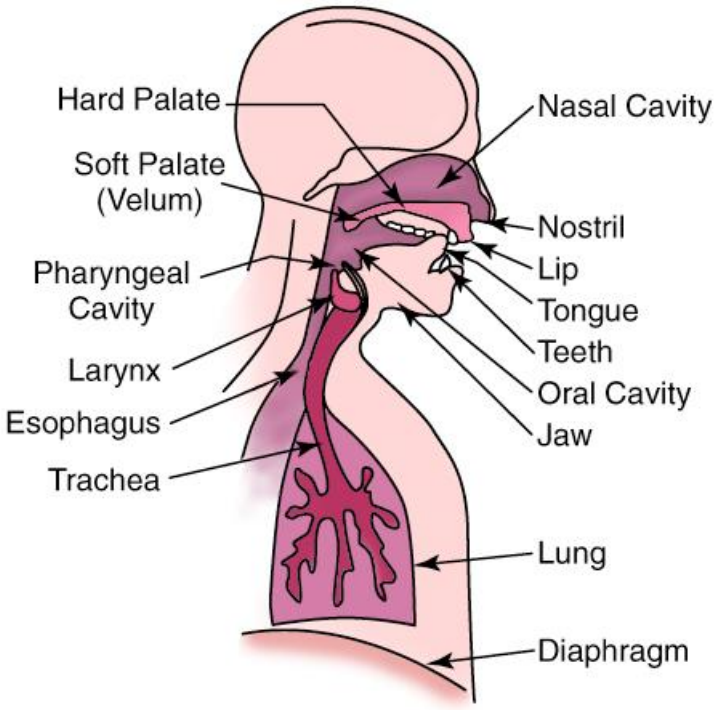

Figure 2.1. Schematics of Human Speech Production Mechanism [19]

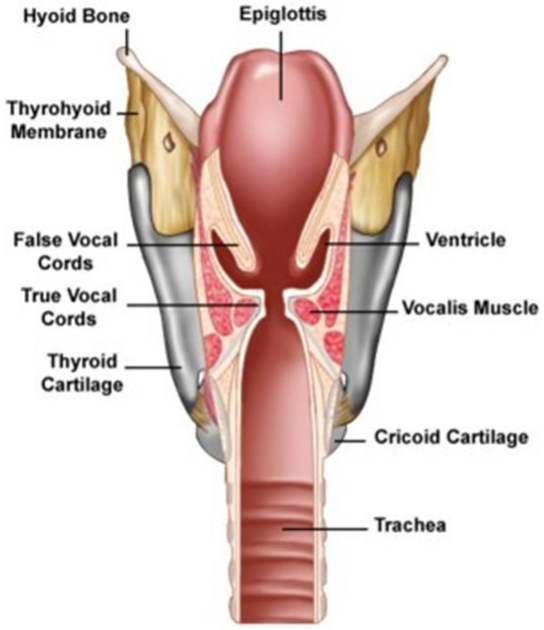

Figure 2.2. Cutaway View of the Larynx [20]

and hence, for "voicing". Voicing is resulted in response to the sub-glottal air pressure changes according to the Bernoulli principle and the concept of energy conservation. The vocal cords vibrate causing the glottis to open and close periodically, and these rapid interrupts in the air passing through the glottis generate quasi-periodic pulses

and produce voiced phonemes like vowels or voiced consonants. With stop consonants, the vocal cords show a rapid change from a completely closed position to a fully open position. For the unvoiced sounds, on the other hand, the glottis remains open and a noise-like excitation is driven by the air passing through a constriction somewhere along the vocal tract.

Table 2.1. A Rough Classification of Turkish Phonemes

	Vowels	**Consonants**			
	All:	Fricatives:	Plosives:	Nasals:	Liquids:
VOICED	/a/, /e/, /ı/, /i/, /o/, /ö/, /u/, /ü/	/c/, /j/, /v/, /z/	/b/, /d/, /g/	/m/, /n/	/ğ/, /l/, /r/, /y/
UNVOICED	-	/ç/, /f/, /h/, /s/, /ş/	/t/, /k/, /p/	-	-

In case of voiced sounds, in standstill the vocal folds are situated close together and the glottis is almost closed. When phonation begins, the increasing air pressure pushes the vocal folds apart, and thus, the glottal hole begins to open. As the air flow keeps opening the glottis wider, glottal volume velocity and the kinetic energy rise, while air pressure and its potential energy decrease. So, at some point, the pressure, reaching its minimum, makes the vocal folds begin to close. When the vocal folds are closed again, air pressure starts building up and the cycle is repeated in the same way as long as phonation remains.

The rate at which the vocal folds open successively is called the fundamental frequency[1] F_0 $(1/T_0)$. The fundamental frequency depends on the sub-glottal air volume as well as on the size, rigidity and tension of the vocal folds at that moment. Men generally having a lower fundamental frequency than that of women is thus reasonable, since the average size of the vocal folds is larger in males.

It is also to be noted that the glottal pulse is skewed to the right, i.e. the "opening phase" is slower than the "closing phase". The stored kinetic energy during opening

[1]There is this term "pitch" which refers to the perceived fundamental frequency. Nevertheless, both terms are very frequently used interchangeably in literature and so throughout this thesis.

phase turns into an elastic restoring force, which together with Bernoulli force acts to close the vocal folds abruptly [21].

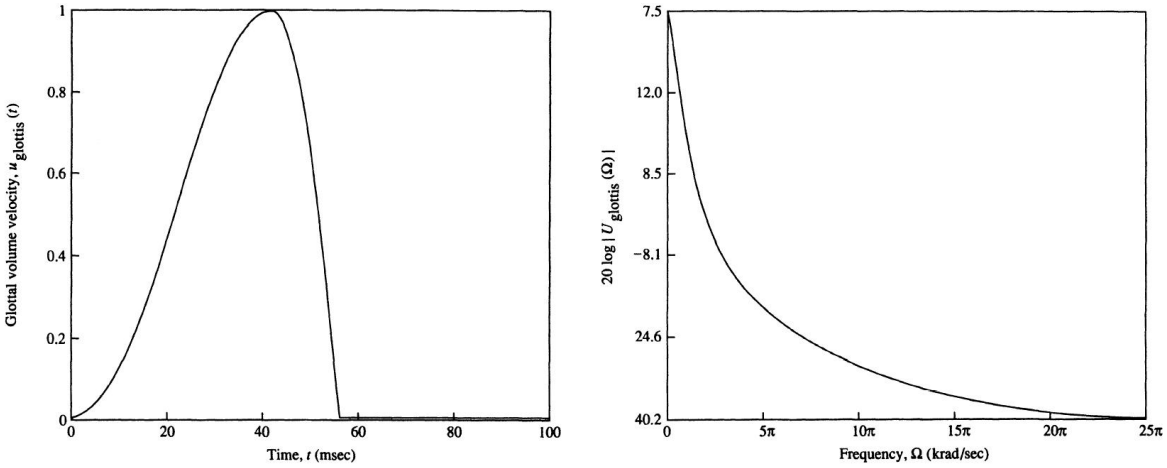

Figure 2.3. Time-Domain Waveform of a Glottal Pulse and Its Magnitude Spectrum [21]

The glottal muscular tensions affect the spectral characteristics of the glottal waveform. The lowest frequency components are mainly determined by the glottal pulse skewness, while the higher frequencies and the spectral tilt are related to the abruptness of the glottal closure [22-26]. In addition to the fundamental frequency, these factors have an impact on the voice source quality in terms of breathiness, harshness, creakiness, etc. [22]. Voiced phonation also includes an aspiration noise, which is important during breathy speech and whispers.

2.2. Speech Modeling

From the engineering point of view, researches on understanding the speech production rely on several models and mathematical representations of the mechanisms mentioned in Section 2.1. Two prominent techniques are the Sinusoidal Modeling and the Source-Filter Modeling.

2.2.1. The Sinusoidal Model

First introduced (McAulay and Quatieri) in the 1980's, sinusoidal models treat speech signals as the sum of a number of sinusoids each with their own amplitudes,

frequencies and phases varying over time. The representation of the speech waveform in this model is given as

$$s_n = \sum_{m=1}^{M} A_m(t) \cos\left[\int_0^t \omega_m(\tau)d\tau + \phi_m(0)\right] , \tag{2.1}$$

where s_n is the n-th sample of speech signal and $A_m(t)$, $\phi_m(t)$ and $\omega_m(\tau)$ denote the belonging amplitude, the phase and the instantaneous frequency of the m-th sinusoidal component respectively. The parameters of the model are found by short frame peak-peaking the DFT spectrum. Then a nearest neighbour matching algorithm and interpolation are employed to relate the frequencies in the adjacent frames. Re-synthesis is done by substituting the achieved parameters into Equation 2.1.

New sinusoidal models have also appeared due to some modifications made to the original. One technique, for instance, is the *Analysis-by-Synthesis/Overlap-Add Model (ABS/OLA)* by George and Smith [27], which outputs very good results in altering musical voice. It employs an iterative analysis-by-synthesis procedure for parameter estimation and eliminates the need for parameter tracking by the use of the overlap-add procedure. Another popular approach, the *Harmonic Plus Noise Model (HNM)* by Stylianou [28, 29] makes use of harmonic relations and the quasi-periodic character of the speech signal. Here the signal is assumed to consist of a *harmonic* and a *noise* part that are treated separately in the analysis stage. Synthesis is again performed in an overlap-add fashion.

2.2.2. The Source-Filter Model

Although sinusoidal models outperform other methods in applications that require high-quality speech synthesis, thinking of speech production as an acoustic filtering process secures greater flexibility for modifying the features of the speech signal.

In this approach, pharyngeal, nasal and oral cavities constitute together the acoustic filter whose characteristics are defined by anatomy and by the positions of

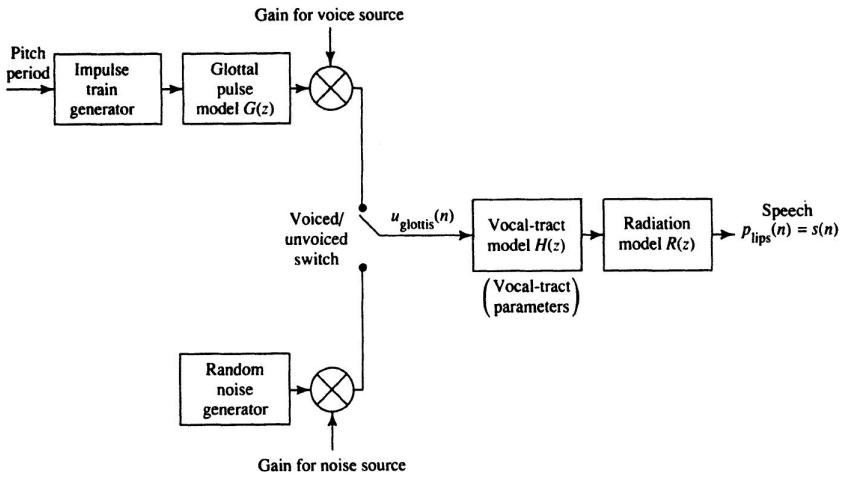

Figure 2.4. Human Speech Production Mechanism in Block Diagram (after Rabiner and Schafer-1978) [21]

the articulators. The input signal to the filter is generated by the vocal cords at the larynx out of the air flow from the lungs (as explained in Section 2.1). The lips and partly the nostrils act as low-impedance loads where speech is radiated in the form of sound pressure waves.

The Source-Filter Model, whose flowchart is given in Figure 2.4, is a more simplified representation based upon this approach.

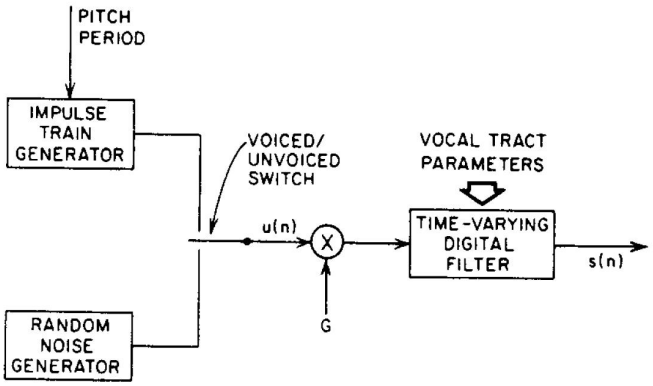

Figure 2.5. Flowchart of the Source-Filter Model [30]

As mentioned above, the excitation of the filter can be voiced, unvoiced (or a mixture of both). Since unvoiced excitation is turbulent and noise-like due to a significant vocal-tract constriction waylaying the airflow, it is simply modeled to be a random noise generator. Voiced excitation, on the contrary, is the quasi-periodic glottal wave

related to the oscillatory movement of the vocal folds. This source type can thus be represented as an impulse train generator, although various parametric models have been proposed in the literature [22, 31] to better describe its special shape. Sometimes an all-pole digital glottal shaping filter $G(z)$ is placed right after the impulse train generator for higher accuracy [21].

According to the Lossless N-Tube Model [21], the acoustic filter (in other words, the vocal tract filter) itself is modeled as a time-varying, p-th order all-pole filter (Equation 2.2) driven by the source excitation

$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \ , \tag{2.2}$$

where G stands for overall gain and a_k are the complex poles of the transfer function. The details of this filter assumption will be covered in the following sections.

Finally, the radiation at the lips (and the nostrils) further shapes the signal by acting as a digital differentiator of Equation 2.3.

$$R(z) \approx 1 - z^{-1} \tag{2.3}$$

Obviously, this filter has a single zero at $z_0 = 1$. It is normally recommended to decouple R(z) with G(z), so that the zero cancels out and the all-pole nature of the overall model is preserved. The resulting sound pressure at the end of the whole system is then the speech signal s(n).

2.2.3. Estimating the Filter Parameters

2.2.3.1. Linear Prediction Coefficients (LPCs). Linear predictive technique is one of the most widely used methods for filter parameterization of the Source-Filter concept. In this thesis, Linear Prediction (LP) is preferred not only due to this reason, but also because of its compeering use in speech modification process which will follow directly.

For simplicity, glottal shaping filter, vocal tract filter and lip radiation in Figure 2.4 can be merged into a single time-varying filter whose steady-state function is given as

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{2.4}$$

Such a system dictates the simple difference equation

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n) \ , \tag{2.5}$$

where $u(n)$ is the source excitation (voiced/unvoiced), G is the gain parameter and a_k are the LP coefficients.

Equation 2.5 apparently suggests that the sample $s(n)$ of a speech signal at a moment is related to the linearly weighted summation of its past values such that it can be modeled by a pth-order linear predictor with prediction coefficients, α_k.

$$s(n) \approx \bar{s}(n) = \sum_{k=1}^{p} \alpha_k s(n-k) \tag{2.6}$$

Then, the prediction error (often also referred to as the LP residual) in regard to Equation 2.6 is defined as

$$e(n) = s(n) - \bar{s}(n) = s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \tag{2.7}$$

A comparison of Equation 2.7 with Equation 2.5 reveals the obvious fact that the speech signal fits into the LP equation perfectly by choosing $\alpha_k = a_k$ and $e(n) = Gu(n)$. Hence, the transfer function of the prediction error, the prediction error filter, becomes

the inverse filter of $H(z)$:

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} \tag{2.8}$$

This observation makes the use of linear predictive technique quite justifiable: For the linear predictive model to better approximate the speech signal, the error $e(n)$ in Equation 2.7 must be minimum. As a matter of fact, for voiced speech, $u(n)$, and hence $e(n)$, is a train of impulses meaning that it is mostly very small. Similarly, for unvoiced phonemes, the noise-like excitation has a relatively low energy.

The next step is to determine the Linear Prediction Coefficients (LPC). Having given the clue right above, this is done by minimizing the prediction error (mean-squared prediction error) on a frame-by-frame basis due the time-varying nature of the speech signal.

$$\begin{aligned} E &= \sum_n e^2(n) = \sum_n \left[s(n) - \bar{s}(n)\right]^2 \\ &= \sum_n \left[s(n) - \sum_{k=1}^{p} \alpha_k s(n-k)\right]^2 \\ &= \sum_n s^2(n) - \sum_{k=1}^{p} \alpha_k \sum_n s(n)s(n-k) \end{aligned} \tag{2.9}$$

There are two common approaches, autocorrelation and covariance methods, to solve the resulting set of normal equations. The simplified all-pole filter model that the linear predictive method assumes is a good estimation for non-nasal sounds, but for nasal and fricative sounds, additional zeros are introduced in the original acoustic theory. Nevertheless, if the order p in the linear predictor is chosen sufficiently large, then it still outputs satisfactory results overall.

The ease and efficiency in computing the model parameters is the major strength of LP analysis. Its success to represent the spectral envelope of the speech signal is also

advantageous in tasks like voice conversion where it is useful to manage the spectral properties with few parameters. Nonetheless, its poor interpolation and quantization properties are the main disadvantages of the LP system; small changes in LPCs affect the frequency response of the vocal tract filter. Thus, Line Spectral Frequencies (LSF) derived from the LPCs are sometimes preferred to overcome these drawbacks [22, 32, 33].

2.2.3.2. Line Spectral Frequencies (LSFs). The set of Line Spectral Frequencies (LSFs) or Line Spectrum Pair (LSP) contains the same information as LPCs and was introduced in the 1980's as an alternative to them.

LSFs are calculated by developing a symmetric and an anti-symmetric polynomial out of the pth-order inverse filter, $A(z)$, in Equation 2.8 such that

$$A(z) = \frac{1}{2}(P(z) + Q(z)) \tag{2.10}$$

with two (p+1)th-order polynomials

$$P(z) = A(z) - z^{-(p+1)}A(z^{-1}) \tag{2.11}$$

$$Q(z) = A(z) + z^{-(p+1)}A(z^{-1}) \tag{2.12}$$

The polynomials $P(z)$ and $Q(z)$ correspond to the Lossless Tube Model of the vocal tract with a closed and open glottis, respectively [21], and the roots of these polynomials are the LSF parameters.

After having been introduced as a means of representing the vocal tract features by Arslan et al. [32], one of the reasons that the use of LSFs has become popular is the ease in stability check. The vocal tract filter response is stable iff the roots of $P(z)$

and $Q(z)$ lie on the unit circle alternate in order [34]; and the conditions that (i) all the roots are indeed located on the unit circle and (ii) that they are interleaved have been proven to be satisfied [33, 35].

Computed by several different methods like iterative search along the unit circle by taking advantage of the interleaving or Newton-Raphson approximation [34], LSFs also possess other useful properties such as their good interpolation and quantization possibilities, only the knowledge of phase or angle being sufficient to describe them (since the magnitude is unity) and their relation to formant locations and bandwidths (a formant being likely to occur between two closely placed adjacent LSF values).

3. SPECIAL CASE: DYSPHONIC SPEECH

Since this study primarily focuses on restoring dysphonic speech, or alaryngeal speech in particular, it is thought to be necessary to give the essential background about the term "dysphonia". One speaks of alaryngeal speech in cases where the larynx cannot accomplish its vocal functions and speech is created using sources other than the vocal folds. These phenomena may be resulted from some defect in the glottal area (e.g. vocal fold paralysis) or most probably from complete removal of larynx (because of larynx cancer). This chapter investigates laryngeal disorders from a pathological point of view after a more detailed description of the larynx. Then, alternative methods to substitute laryngeal phonation are introduced; their differences from the normal speech and from each other are discussed.

3.1. A Closer Look to Larynx

The larynx, also known as "voice box" or "Adam's apple", is a complex organ situated in the neck of mammals, which is responsible for protection of the lungs and for production of speech. In humans, it is placed at the junction of pharyngeal tract, trachea and esophagus, connecting the hypopharynx (the inferior part of the pharynx) with the trachea. The primary biological function of the larynx is to guard the airway from food, fluids or other foreign objects by closing the glottis during swallowing or coughing. While swallowing, the epiglottis covers the larynx and passes the food through the esophagus to the stomach. Voice generation is a not so crucial but still significant secondary function, essential for oral communication. The larynx houses the vocal folds, which alter pitch and voicing as described in Section 2.

Often divided into three sections (supraglottis (the area above the vocal cords), glottis (the area containing the vocal cords), and subglottis (the area below the vocal cords)), the larynx is a highly complicated structure of four basic anatomic structures: a cartilaginous skeleton, muscles, innervations and a mucosal lining.

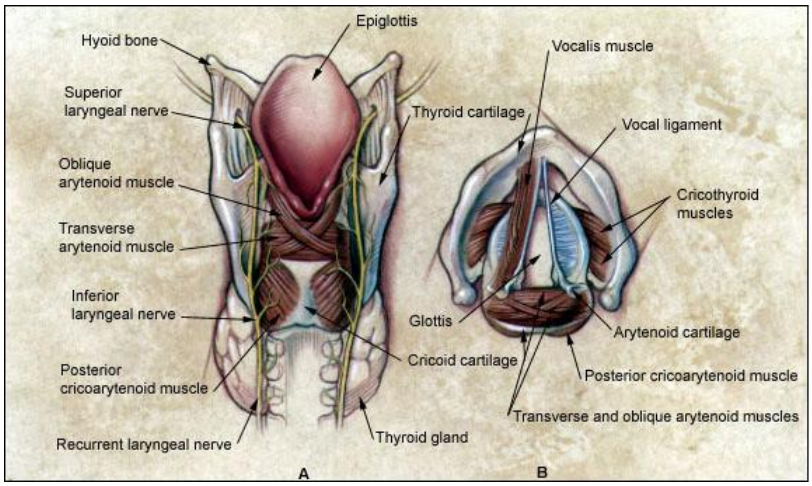

Figure 3.1. Cross-sections of the Larynx [36]

The cartilaginous skeleton contains the vocal cords and is formed by nine cartilages, three single (thyroid, cricoid, and epiglottic) and three paired (arytenoid, corniculate, and cuneiform). Lined with *the mucous membrane*, the cartilages are connected to each other by muscles and ligaments. *The extrinsic muscles* connect the cartilages to other structures of the neck and head. *The intrinsic muscles* serve to change the shape, position and tension of the vocal folds and the glottal volume. The internal branch of *the superior laryngeal nerve* acts as sensory innervation to the glottis whereas its external branch stimulates the cricothyroid muscle, the intrinsic muscle which lengthens and stretches the vocal folds. Sensory innervation to the subglottis and overall motor stimulus is by *the recurrent laryngeal nerve*. Disorders of the external laryngeal nerve results in inability to tighten the vocal cords, and hence in weakened phonation. Damage in the recurrent laryngeal nerves, on the other hand, causes hoarse speech [37-39].

The vocal component of the larynx possesses two mucosal fold pairs: false vocal cords (vestibular folds) and (true) vocal cords. Despite to a few exceptions, the false vocal folds do not play a supreme role in voicing, they are rather responsible for resonance. The vocal folds form a V-shape in the horizontal plane. During phonation they get close by bringing together the arytenoid cartilages and are set into vibration by the expelled air. The muscles attached to these cartilages regulate the degree of glottal opening whilst the backward-forward movement of the thyroid cartilage on the cricoids

cartilage controls the vocal fold length and tension. This, in turn, causes the pitch to rise or fall. Vocalis muscles help to change the tone by increasing the thickness of the chords [39].

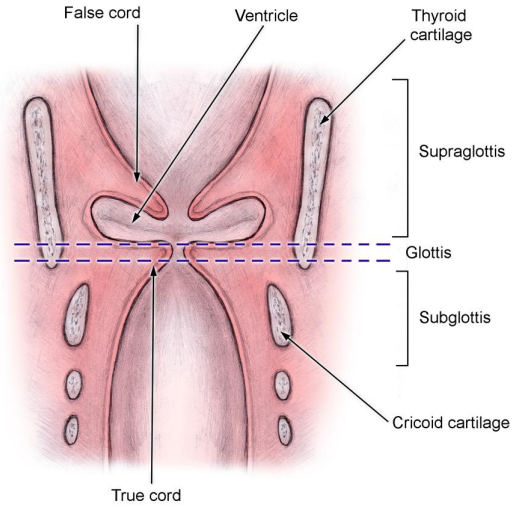

Figure 3.2. Anatomical Regions of the Larynx [40]

3.2. Laryngeal Cancer

There are several disorders that prevent the larynx from functioning properly. These range from a simple inflammation caused by the common cold to polyps and nodules, which affect the vocal fold vibration negatively.

All those resulting in dysphonia to different degrees, a much more serious disease regarding the larynx is the laryngeal cancer. Laryngeal cancer constitutes 2% of all cancer types that occur in adults. Though the rates of female patients seem to have increased in recent years, it occurs about 4 times more commonly in men than in women. Smoking and alcohol use are named as the main risk factors for this disease, but sometimes it also associated with genetic and viral factors as well as radiation and asbestos exposure.

In spite of the fact that laryngeal tumors can develop in any location in the larynx, it originates most commonly in the glottis. Supraglottic and subglottic cancers are less and least frequent, respectively. Symptoms like a chronic sore throat or slowly

developing, persisting hoarse voice are regarded as possible signs for laryngeal cancer.

In modern medicine, there are several procedures for treatment varying with the location, type, and stage of the tumor. These operations include radiotherapy, chemotherapy, laser surgery (vaporization or cutting out the tumor, sometimes some parts of one vocal cord, using a high-intensity laser), pharyngectomy (removing all or part of the pharynx), cordectomy (removing the vocal cords completely or partially in very limited or superficial glottal cancer types), laryngectomy or some of them in combination. Tumors at an early stage can be treated with the former methods. However, larger tumors and the recurrence of the disease mostly make laryngectomy unpreventable [41, 42].

3.3. Laryngectomy

Laryngectomy is the surgery where all or part of the larynx is removed. Laryngectomies are occasionally accompanied by other treatments, such as radiation or chemotherapy. If the tumor is small, it may be cured with a *partial laryngectomy* where only part of the voice apparatus is taken out. Differing in procedure depending on the affected area, it always targets the same goal of removing the cancer entirely while preserving the functionalities as much as possible. In the event of a partial laryngectomy, it is usually managed for some speech to remain (or even retain the normal speech in some supraglottic cancers), though in most of the cases the resulting speech is very hoarse after the operation.

Contrariwise, cancers at later stages (e.g., Stage III and IV laryngeal and hypopharyngeal cancers) necessitate complete removal of the larynx or even other neighbour structures in the throat and the neck. As a result of the *total laryngectomy*, all anatomical functions of the larynx are lost. During the surgery the surgeon needs to perform a tracheotomy, separation of the pharynx and the trachea. The trachea is brought up to the skin by making an artificial opening called a "stoma" in front of the neck to secure breathing and protection of the lungs. The junction between the pharynx and the esophagus is kept, so the laryngectomees can eat normally except a

few encounter with some eating issues in the healing period [41, 42].

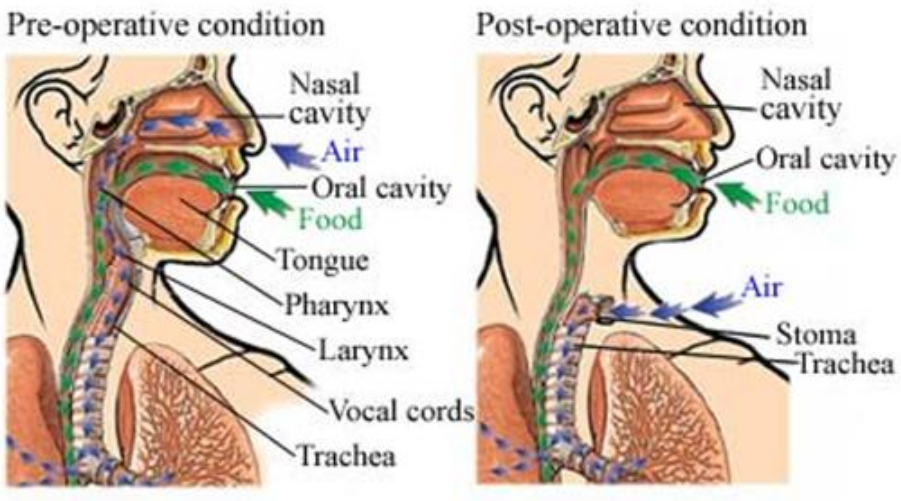

Figure 3.3. Surgical Removal of the Larynx

Unfortunately, the loss of (normal) speech is maybe the most disturbing consequence of total laryngectomy, both pathologically and psychologically. Hence, speech rehabilitation becomes the central point of post-laryngectomy period. There are indeed some methods of vocalization after a total laryngectomy (described hereafter), still one can hardly speak of a complete speech restoration for alaryngeal speech having a poor quality and an inadequate level of intelligibility.

3.4. Speaking after Laryngectomy

In partial laryngectomy, the surgeon is usually able to leave some parts of the larynx behind which, in turn, allows the speech to remain to a certain degree: The remaining tissues, cartilages or walls of the larynx make vibration possible in the absence of the vocal folds. Sometimes only one vocal fold is taken out. In case of supraglottic laryngectomies, it is even possible to keep the normal speech since the surgery only involves the removal of the portion above the vocal folds, not the vocal folds themselves. Nonetheless, of course, this is a rare case. Partial laryngectomies typically result in hoarse, wheezy, noise-like speech low in amplitude.

After total laryngectomies, on the contrary, normal speech using the vocal apparatus is no longer possible since the entire larynx has been removed. So, there is

a need for alternate means of vocalizing to sustain oral communication in some sense and, thus, laryngectomees also start a speech rehabilitation phase with a speech therapist or speech pathologist following the surgery and the aftercare. These alternatives (of which the choice depends on several factors like the patient's age, medical state or abilities) are as follows :

3.4.1. Esophageal Speech

Because of the separated trachea from the pharynx, a total laryngectomee can no longer speak using the air expelled from the lungs through the mouth. Instead, they can learn to swallow air down into the esophagus and create a belch-like sound by releasing the air with simultaneously articulating the words. Since it is the walls of the upper esophagus and not the (missing) larynx that vibrates, the esophageal speech is described as "speaking by eructation".

The resulting speech is often harsh and of low intensity. Its pitch is very low (between 50 Hz and 100 Hz) due to the larger area oscillating (upper segment of the esophagus) than that (vocal folds) of normal speech. There is also a direct relationship between the pitch and speech loudness: low-pitched sounds are uttered with lower energy and vice versa; the latter of which is produced with more effort.

Esophageal speech is the most natural way for post-laryngectomy speech production; it does not require any other surgery, prosthesis or any additional battery-driven device. Notwithstanding, this technique is a skill that usually takes several months to gather and not all laryngectomees manage to master.

3.4.2. Tracheoesophageal Speech

This method implies the surgical insertion of a voice prosthesis (Tracheoesophageal Prosthesis or Puncture, TEP) either during or 10-14 days after the initial laryngectomy. Placed into an artificially opened duct between the trachea and the esophagus, the prosthesis is a one-way air valve which enables the exhaled air to be directed through the

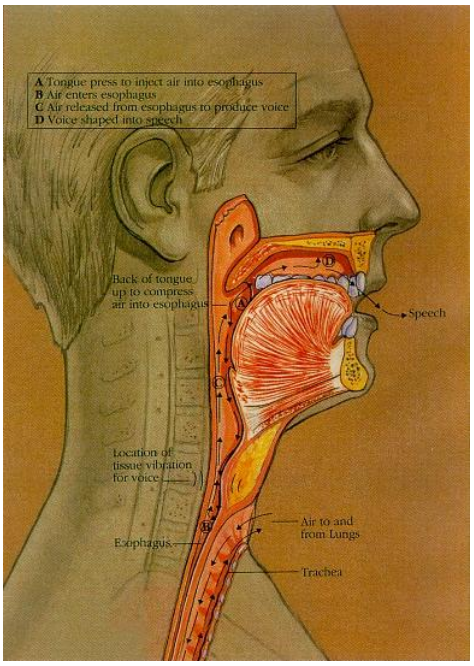

Figure 3.4. Esophageal Speech [43]

esophagus. The pulmonary air is forced through the valve into the esophagus and then to the mouth by occluding the stoma during exhalation and this air movement sets the esophageal walls into vibration. Blockage of the stoma can either be done manually with a finger (e.g. thumb) or with a hands-free tracheostoma breathing valve that automatically blocks the air for speaking. The one-way shunt serves for the prevention of swallowed food, liquids or saliva from entering into the lungs while allowing the air to pass through.

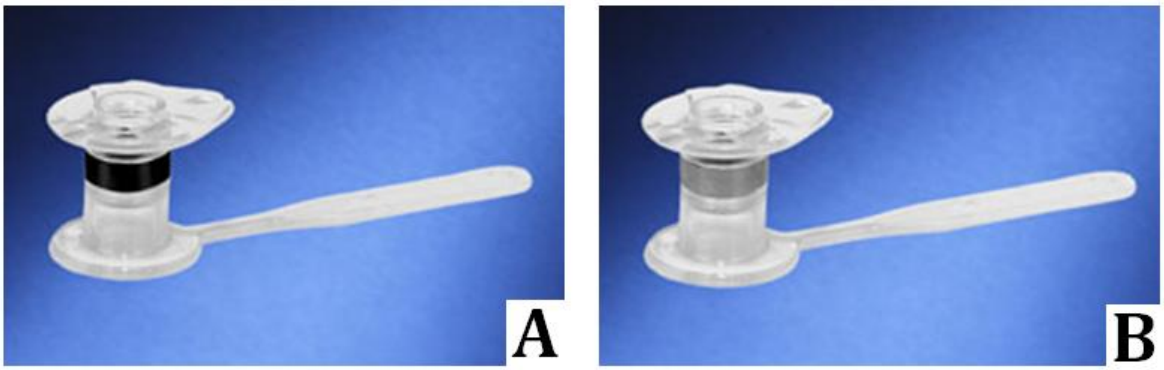

Figure 3.5. Soft Valve Assembly (A) and Hard Valve Assembly (B) Prostheses (samples from InHealt Technologies) [44]

Tracheoesophageal speech is often described as the most natural-sounding of all speech restoration methods in terms of fluency, quality and intelligibility.

TEP technique was introduced by Singer and Blom in 1979 [45]. Its biggest advantage is that a fluent and a relative good-quality speech can be achieved in very short time after the prosthesis surgery. Thus, especially with the advent of new technologies in the past ten years, it has become increasingly appealing for patients.

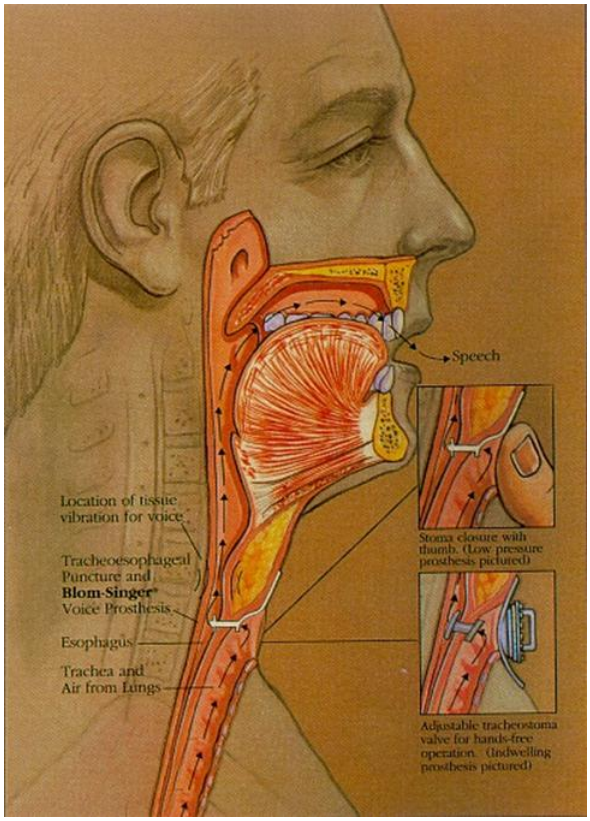

Figure 3.6. Tracheoesophageal Speech [46]

3.4.3. Electrolarynx (EL or Artificial Larynx)

Electrolarynx (EL) is a battery powered, hand-held device which possesses a low-frequency oscillator. In commercial use, there are two types of the electronic larynx: One model has an electromechanically vibrating diaphragm that, when placed against the neck or the floor of the mouth, transmits the vibrations at (more than) one frequency through the skin to be formed into words with the help of the vocal tract and the rest of the articulators. With the second (intraoral) type, the vibrations are radiated directly into the mouth via a small plastic tube connected to EL. This is useful in cases where efficient transmission of the acoustic energy cannot be achieved due to inelastic tissues, irregular tissue contours, scars, etc. It is also suitable for patients

with a short neck who find pressing the electrolarynx too painful.

Most prominent offerings of EL are the ease of use and the rapid gathering of vocalization with almost no training. However, speech produced by an electrolarynx sounds monotonous, buzzy and robotic. If the second type is used, it is also possible that the pronunciation of some sounds ('g', 'k', 't' and 'd') may interfere with the vibrations of the tube which further reduces the intelligibility. Additionally, the user has to pay attention to the positioning of EL; it must always be placed at the same position correctly to secure a good transfer of vibrations. The occupation of at least one hand each time is another practical disadvantage. (An intra-oral artificial larynx that can be worn is proposed pointing out this problem.) For these reasons, this final method is not commonly preferred over the previous ones, especially if the quality of communication is the main concern. Patients generally appeal to electrolarynx in order to overcome the speech rehabilitation period easily, or it is used in cases when esophageal speech and TEP are somehow not possible.

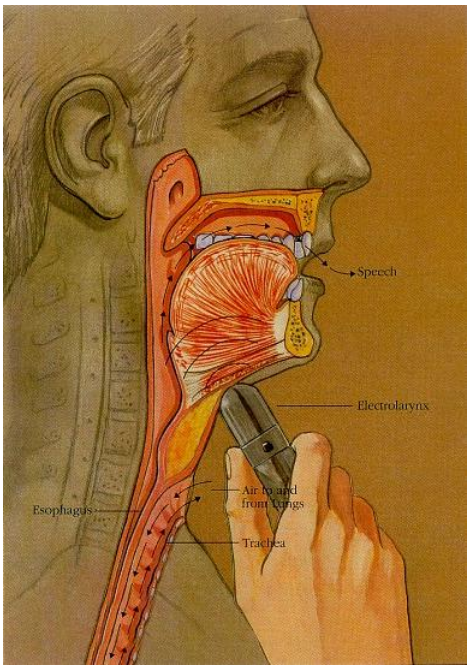

Figure 3.7. Electrolarynx [47]

4. VOICE CONVERSION FOR DYSPHONIC SPEECH REPAIR

In this chapter, the proposed dysphonic speech repair techniques are formulated, which are two different interpretations of the same system based on Voice Conversion technology [32, 33, 48] and the Source-Filter Model for speech signals. Both techniques follow the general two-staged framework of voice conversion (see Fig. 4.1) where the basic distinction between these two algorithms is that one works solely *in time domain* whilst the latter relies on modifications *in frequency domain.*

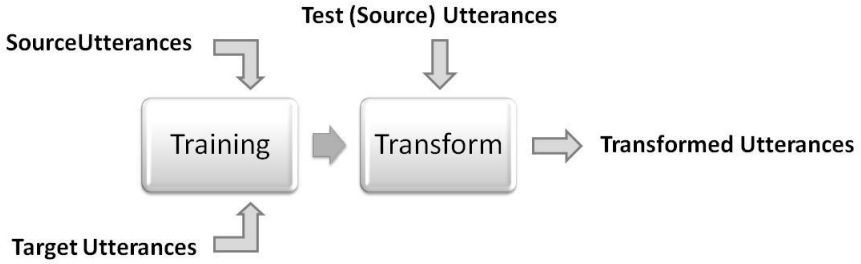

Figure 4.1. General Framework of Voice Conversion in Rough Scale

First, a parallel set of phrases is uttered by both dysphonic (source) and laryngeal (target) speakers as training material. After alignment, the model parameters are extracted for both speaker types to generate source and target codebooks. At the transformation stage, these pieces of information are used for transforming the source and the filter parameters.

4.1. Training

Training phase is responsible for obtaining model parameters which are required for the second stage, for the modification of dysphonic speech. Here, a set of identical utterances are recorded from source and target speakers. Identical source and target utterances in the training corpora are paired, and each recording pair is passed through the following successive steps that, at the end, result in distinct source and target codebooks:

- Labeling: First, source and target files are labeled such that phoneme boundaries in each waveform are marked. Generating the label files automatically can hardly be considered as an option since the poor quality of the dysphonic speech makes recognition outputs in acceptable levels almost impossible.

- Short-term analysis: 512-point sliding Hamming windows are taken for feature analysis. In a source-target pair, the skip rate is fixed (for instance, 5 ms) for source signal, but it is automatically adjusted by a phoneme-duration scaling factor for target signal so that equal number of parameters are extracted for each phoneme in the signal pair.

- Feature extraction: Features that are to be gathered for each frame via short-term windowing correspond to the distinctive parameters in Source-Filter Model (see Figure 2.5); linear prediction coefficients representing the vocal tract parameters of the time-varying filter, gain factor G and fundamental frequency, F_0 (or frame pitch value). F_0 is set to be zero for all source frames as there is no pitch present in the dysphonic speech. LPCs are then converted into and stored as line spectral frequencies due to reasons mentioned in Chapter 2.

In result, two codebooks are generated for the source and the target speaker separately with three types of entries: line spectral frequencies (LSFs), frame gain and frame F_0 values. The same procedure is applied to all of the source-target pairs so that every resulting source and target codebooks are appended one after another to create two large batch codebooks of equal size and matching entries which cover the whole training material.

4.2. Transformation

The transformation stage focuses on the vocal tract filter and source excitation components individually. As the reader may have noticed, there has been no dissimilarity between the two techniques developed for this study in terms of the training of the codebooks. However, these two approaches make use of different methodologies *in the way they model the source component* while reconstructing the dysphonic speech

signal.

The first method strictly adheres the conventional time-domain interpretation of the Source-Filter model where the excitation is added in time domain as the residual signal for unvoiced frames (instead of random noise for the sake of naturalness) and as a combination of the residual signal and impulse train with the fundamental frequency F_0 (of the corresponding target frame) for voiced frames.

The second approach operates in frequency domain from both the source and filter point of view to keep better track of spectral properties. The vocal filter is fed with the extracted glottal excitation spectrum of the target frame as the residual signal for voiced frames; for unvoiced frames, frame residual spectrum of the test data is used as excitation.

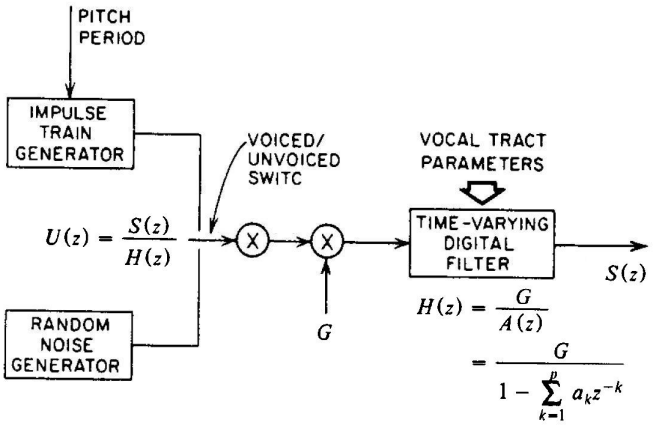

Figure 4.2. Frequency-Domain Version of the Source-Filter Model [30]

This approach makes itself justifiable when one thinks of the frequency-domain version of the Source-Filter Model in Figure 4.2. Here, the excitation signal can be interpreted as the signal spectrum divided by the vocal tract response such that

$$U(z) \approx \frac{S(z)}{H(z)} \tag{4.1}$$

Accordingly, in the frequency-domain technique, this resulting spectrum (computed once for a voiced target frame in the training material) is also given to the transformation function additionally to characterize the glottal excitation in voiced frames

during the transformation. For unvoiced frames, excitation is represented by the frame residual spectrum of the test data itself.

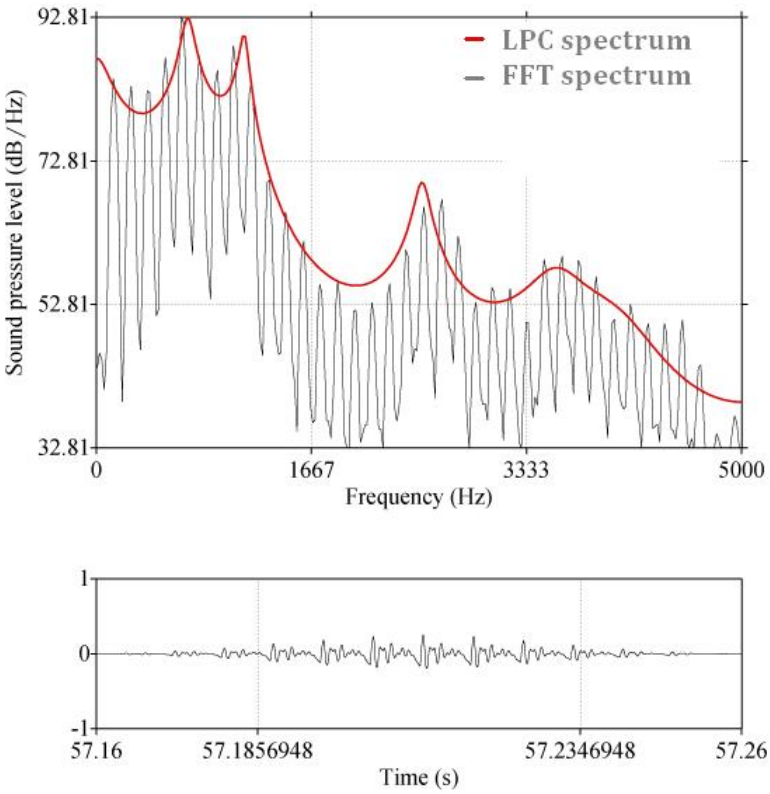

Figure 4.3. LPC and FFT Spectra

As for the vocal tract filter, LSFs of the dysphonic speech to be transformed are again computed frame-by-frame with the same frame length and fixed skip rate as in the training. They are used for searching the closest matching LSF set in the source codebook having the smallest mean-squared error (MSE) distance and are then replaced with the target speech counterparts. This LSF set is then converted back to LPCs to be joined with an appropriate (voiced or unvoiced) source excitation signal either additively in time domain as in the time-domain approach or multiplicatively in frequency domain (the frequency-domain approach). The voiced/unvoiced decision is made by setting a threshold (30 kHz) for the target frame F_0.

5. TESTS AND EVALUATIONS

This chapter presents the test results of the developed enhancement schemes which are applied to two different test cases. In Section 5.1, the reader is provided with the speech material used for evaluations. A set of subjective listening tests are conducted to establish their performances in terms of naturalness and intelligibility. Finally, a discussion of the results is given in the last section.

5.1. Speech Corpora

Five different types of dysphonic speakers were asked to record speech utterances from data sets CUM1 and/or CUM2 given in Appendix A. Patient No. 1-4 were patients in Ear, Nose and Throat (ENT) Clinic of Istanbul Faculty of Medicine and they participated in this project under supervision of a speech pathologist. The fifth speaker, Patient No. 5, volunteered individually and provided his pre- and post-recordings for this study.

5.1.1. Data Collection Setup

Case 1: Patient No.1 is a larynx cancer patient who has undergone a total laryngectomy operation in 2009. He has not practiced any of the voice rehabilitation methods and is not using any assistive vocalization device. The process of recording was very difficult and painful to this patient, so only the first 20 CUM1 utterances were managed to be recorded.

Case 2: Patient No.2 had a regional tumor at the glottal area which has been removed in 2010. He has not fully lost his phonation due to the remaining tissues. The whole data set CUM1 was recorded, of which five utterances (41-45) are used as test data.

Case 3: Patient No.3 is total laryngectomee who has been operated in 2005. He

is using an electrolarynx to overcome his lack of phonation. Since he found reading out all utterances very tiresome, 45 phrases (1-40 as training, 41-45 as test material) CUM1 data set were recorded.

Case 4: Patient No.4 has also undergone a total laryngectomy surgery in 2005. He is a practiced esophageal speaker who masters this speech production method. All items in CUM1 were recorded where utterances 41-45 are used as test data.

For these four cases, the recordings of the dysphonic speech material took place in a private room in the ENT Clinic of Istanbul Faculty of Medicine. A normal voiced, professional male speaker recorded the same utterances that served as parallel target stimuli in an acoustically isolated audio recording room at Sestek Inc. All dysphonic and normal speech recordings were carried out at a sampling frequency of 16 kHz.

Case 5: This final case differs from the preceding test cases in terms of data set content and recording environment. Patient No.5 has a larynx cancer spread along the glottis. Although he does not have his larynx removed, his voice is almost fully lost due to the disease. Here, both dysphonic and normal data was provided (at 44 kHz) by the speaker himself with normal speech files already present and dysphonic utterances taken in the audio recording room at Sestek Inc. 60 utterances from CUM2 are used for training, last 5 utterances serve for test purposes.

During recordings, all patients were asked to speak at a comfortable loudness level and speech rate, and they were able to pause or repeat the utterance as many times as they needed.

5.2. Subjective Tests

The described speech database is used for testing the repair algorithms presented in this work. The system is trained with parallel dysphonic and normal corpora and the material reserved for tests (utterances 41-45 in CUM1 by Patients No. 1-4 and utterances 61-65 in CUM2 by Patient No. 5) is used to perform comparative analysis

Table 5.1. Information on Test Subjects

Patient Information			Medical Records		Speech Material
No.	Age	Sex	Type of Disorder	Date of Surgery	
1	46	M	Total laryngectomee No phonation	April 2009	CUM1 data set first 20 utterances
2	80	M	Right vertical hemilaryngectomee Husky speech	March 2010	CUM1 data set 100 utterances
3	74	M	Total laryngectomee Electrolarynx (Robotic speech)	February 2005	CUM1 data set 45 utterances
4	55	M	Total laryngectomee Esophageal speech (Belching-like speech)	September 2005	CUM1 data set 100 utterances
5	56	M	Laryngeal cancer patient Hoarse and rough speech	No surgery	CUM2 data set 65 utterances

of reconstructed speech by both time and frequency domain approaches against its dysphonic counterpart. Subjective testing was also applied to evaluate their performance. Five listeners were asked to vote the reconstructed speech in terms of the naturalness from 1 to 5 (5 is the best).

Unfortunately, speech produced by Patient No. 1 was proven to be not suitable to work with, because no parameters can be extracted due to the lack of phonation. The average results of MOS (Mean Opinion Score) tests for **Case 2** and **Case 5** are presented below:

Table 5.2. Subjective listening test results for Test Case 2 (Naturalness)

	Time Domain Approach	**Frequency Domain Approach**
Dysphonic Speech	3.5	3.5
Reconstructed Speech	2.0	2.6

Table 5.3. Subjective listening test results for Test Case 5 (Naturalness)

	Time Domain Approach	**Frequency Domain Approach**
Dysphonic Speech	3.8	3.8
Reconstructed Speech	1.5	1.8

Together with the resemblance of the transformed speech to its normal equivalent (how natural it sounds), the quality of speech is also characterized by how understandably it can be conveyed. The latter aspect actually plays a more important role in

ensuring a healthy communication. Thus, having examined the performances of two proposed methods from the naturalness point of view, additional MOS tests were conducted to observe the contribution of the speech reconstruction to the intelligibility. Based on the findings in the above tests, only the frequency domain approach is taken into evaluation.

Here, the listeners (10 each) were requested to listen to five test utterances (including source and transformed speech files) for each test case and, this time, to rank the intelligibility only. If they fully understood what was said in the speech file, they should grade it as '5'; the files where absolutely no information could be extracted were to be rated a '1'. Speech with lower intelligibility should be ranked in-between according to the level of being understandable. In addition, the test files were inputted to a speech recognition tool (GVZ SR Tool at Sestek Inc.) as another measure of intelligibility. The results are obtained as follows:

5.3. Discussion

Based on the objective measures obtained during development phase, it can be claimed that the proposed schemes succeed in improving the signal spectrum and in adding excitation to some extent into the signals. Subjective listener tests, on the other hand, indicate that speech reconstructed neither in time domain nor in frequency domain schemes are rated to be better than these dysphonic examples, although the method based on frequency domain modification markedly outperforms the time domain approach and increases the success rate. Automatic speech recognition results do also not show any remarkable improvement over the untransformed dysphonic samples, and are thus in accordance with MOS outcomes. The reason why the automatic speech recognition results are near zero in most of the cases not only for reconstructed speech, but also for the dysphonic speech is that the recognizer is not trained for this special type of speech data.

Possible reasons that may have led to the perceptual inefficiency are assumed to be

- uncontrolled recording environment,
- insufficient training data,
- phonetically unbalanced source and target sentences (whilst such a balance helps to obtain successful transformation results),
- poor quality of normal speech samples, and
- (probably, although not been able to be detected) algorithmic weaknesses in the developed code structures.

Table 5.4. Subjective listening test results (Intelligibility)

TEST CASE	**TEST FILES**	**DYSPHONIC**		**RECONSTRUCTED**	
		MOS	SR(%)	MOS	SR(%)
Patient 2	Utterance41 (CUM1)	4.5	25	4.0	75
	Utterance42 (CUM1)	4.4	100	3.7	100
	Utterance43 (CUM1)	4.5	25	4.0	75
	Utterance44 (CUM1)	4.5	25	4.0	75
	Utterance45 (CUM1)	4.5	25	4.0	75
Patient 3	Utterance41 (CUM1)	1.1	0	1.1	0
	Utterance42 (CUM1)	1.3	0	1.2	0
	Utterance43 (CUM1)	1.7	33	2.1	0
	Utterance44 (CUM1)	2.3	0	1.0	0
	Utterance45 (CUM1)	1.6	0	1.1	0
Patient 4	Utterance41 (CUM1)	3.3	0	2.7	50
	Utterance42 (CUM1)	3.0	0	2.7	0
	Utterance43 (CUM1)	3.8	33	3.3	0
	Utterance44 (CUM1)	3.9	0	2.6	0
	Utterance45 (CUM1)	3.7	0	2.5	0
Patient 5	Utterance61 (CUM2)	3.3	N/A	2.6	N/A
	Utterance62 (CUM2)	3.0	0	3.2	N/A
	Utterance63 (CUM2)	3.7	0	3.5	N/A
	Utterance64 (CUM2)	4.1	N/A	3.8	N/A
	Utterance65 (CUM2)	3.3	0	2.8	0

6. CONCLUSIONS

This thesis has addressed the problem of dysphonia from a speech engineering point of view with the aim of reconstructing more intelligible speech from dysphonic samples using speech processing tools. In this study, the Source-Filter Model has been used to model the speech signals. The enhancement mechanism has been based on the notion of voice conversion technology where dysphonic speech frames have been considered as units to be transformed into laryngeal target units by codebook mapping. Listening experiments are conducted to judge the performance of the two speech re-synthesis methods in general and individually. The behavior of the proposed algorithms in different test cases have been tried to be interpreted according to the achieved results.

Based on the findings, the following conclusions can be drawn out of the work presented in this thesis:

- The frequency-domain method outperforms the time-domain approach in dysponia cases considered in this thesis. The first algorithm uses a glottal shape extracted directly from a voiced target frame instead of adding hand-made pulses as the latter does. By doing this, most of the spectral characteristics of the glottal excitation can be preserved and the robotic effects can be eliminated.

- Processing the unvoiced frames adds extra artificialness into the signal, and hence, degrades the overall performance. Thus, better results can be expected by keeping the unvoiced phonemes unprocessed.

- There are cases where a speech re-synthesis is possible using neither of the proposed schemes: Total laryngectomees can produce no voicing and no audio clues computationally detectable. So, the medical rehabilitation solutions are the only chance to sustain phonation for such patients.

- The only case where the frequency-domain algorithm shows a slight improvement is the speech produced with electrolarynx.

- All in all, the codebook mapping and the addition of external glottal excitation alone fails to produce reconstructed speech which has better quality than it's dysphonic equivalent in terms of naturalness and intelligibility.

As a future work, it is planned to revise the proposed algorithms to validate their efficiency and to strengthen their robustness. Additional means of improvement such as statistical modeling, spectral smoothing, excitation estimation or enriching the parameter set can be added to a future system. Diversifying the database and test cases is also considered as a future direction.

APPENDIX A: DATA SET FOR TESTS AND EVALUATIONS

A.1. CUM1

1. amaçlarının ne olduğunu
2. insanların arasındayken
3. sekiz olarak bildirilmektedir
4. bundan sonra da kendisine
5. yörelerinde bulunmaktadır
6. cahillikten başka bir şey değildir
7. erkeklere çalışmalarından bir pay
8. ve bütün bunların üzerinden
9. ne kadar ağaç diktiklerini söyledi
10. kadınlar kendilerine karşı
11. birbirlerinin ayrılıklarını değil
12. bu işin içinde birlikteyiz
13. gerçekten büyük bir insan
14. tarafından karşılanmıştı
15. sonra konuşmaya başladı
16. arasında olması gerekir
17. oysa hiçbir zaman gerçek
18. korkmadığını göstermek için
19. emin olmak için bir daha baktı
20. buna karşı olmayacağını
21. olduğu gibi kabul ediyorlar
22. en önemli özelliklerinden biri buysa
23. ölen bir kadının yanından geliyordu
24. iki türkiyenin böyle bir yeni
25. işlemek istediğini biliyoruz

26. çıkarmaya çalışıyorlardı
27. zat-ı devletleri olabileceğini
28. dudaklarında bir gülümsemeyle
29. düş hakkındaki düşüncelerimi
30. sebada bulunduğunu bildirdi
31. araştırmalarına başlamış
32. belirtilerini kullanması
33. sonunda dedi kendi kendine
34. daima yerine getirilmiştir
35. onlardan birini gönderin
36. yeni bir değerlendirmeyle
37. burada kalmak istiyorum
38. her şeyi ortaya çıkaracak
39. bunun üzerine mustafa kemal
40. eleştirilmesi gereken şey
41. sonra da arkasından istanbula
42. yalnızca bir kez arkadaşlarım
43. gözleri gözlerine değil
44. gerçekleştirebileceklerdi
45. ve benzeri hareketlerini
46. ama bunun için karşısında başka insanlar
47. durumları ve ilişkileri değiştirir
48. olanları yeniden düşündüm
49. bu durumda sivil toplum
50. osmanlı devlet ve milleti
51. ne de bir yardımcı vardır
52. ben de söylemeye başladım
53. yıkılmasına karar verilir
54. ölmek üzere olan bir adama
55. oldukları gibi kaldıkları için
56. dünya ile ilgili bir konudur
57. sadece insan davranışları

58. bu görüşmelerden sonra
59. cumhuriyet gazetesinde
60. insan hakları konusunda da
61. ben onların dışıda kaldım
62. çocukların tanımlamaları
63. herhangi bir durumu alıp
64. bununla birlikte kendimizi
65. eskisinden daha güçlü bir biçimde
66. iç işleri bakanlığından yapılan açıklamada
67. insanın kendine sahip olmadığını
68. elde edilmiş olacağını söyler
69. ekt uygulanan hastalarda
70. içerisinde yapılmaktadır
71. kaybolmasını istemiyorum
72. daha yüksek bir şeylerin
73. babaların yaptığı budur
74. alışkanlık bozuklukları
75. babasının yanına geldiğinde
76. torunlarının bir başarısında
77. belki de daha çok yaşamak
78. ölmesinin nedenlerine gelince
79. bilgilerini verdikleri için değil
80. direklerin altından bağırır
81. elbette doğru söyleyeceğim
82. ellili yılların başlarında
83. tüm zamanlar için geçerli
84. yabancıların kadınlarına
85. bir yandan bunları düşünüyor
86. büyük millet meclisi başkanı
87. bu haberlerin bir kısmında
88. yönetimin gerektiğini ekler
89. büyük değişiklikler olmuştu

90. hep devam etmek zorunda olduğumu
91. kişilerin yalnızlık duyguları
92. der gibi bir halleri vardı
93. ekonomik faaliyetlerin de
94. ailelerini birbirinin başına kaktılar
95. buyruğunda bulunan insanları
96. uzun bir süre kaldırmadı
97. dinin kullanılması da olabilir
98. bu şekilde oluşturulmalıdır
99. ulaşmalarını sağlayacaktır
100. elinden gelen bir hizmeti

A.2. CUM2

1. yangın olur biz yangına gideriz
2. düz ovada keklik gibi sekeriz
3. bu eski bir tulumbacı şarkısı
4. ama a takımının da şiarı
5. hanım ağa
6. telefonlara hem murat sincar bağlanacak
7. hüseyin kocadağın eşi kıymet
8. yıllardır bütün medya mensuplarının peşinden koştuğu
9. bir kare fotoğrafını çekebilmek için
10. bu iddialar tartışılacak
11. ama önce
12. refah millet vekili mustafa bayram kim
13. a takımı düz ovada keklik gibi sekerek gitmiyor yangınlara
14. ya da olaylara
15. bu a takımı adına size ilk merhaba deyişim
16. geçen gün öyle bir yangına gittik ki
17. istanbul trafiğini bir sabaha karşı

18. nejat
19. arkasından da dalga dalga türkiye
20. hatta bütün dünya
21. yeraltı dünyasının
22. hatta uluslararası yeraltı dünyasının
23. sevgilerden şikayet etmeyeceğim
24. haldun hoca duymasın
25. affına sığınarak söylüyorum ayrıca
26. bir ordan vurdular bir buradan yıldızı bizden mahrum
27. ve birlik fırtınası gibi esen hipodrom konserini
28. biraz sonra a takımında neler varmış bir bakalım diyorum
29. gücüm yetmezse alırım ordan üç tane delikanlı
30. ama bundan sonra çok sıkı dost olacağımızı umuyorum
31. günlük güneşlik havalar gelsin artık ülkemize diye
32. refah partisi van milletvekili mustafa bayramla
33. hakkı yok
34. yapıp edip bölücü terör örgütüne fatura ediyorlar
35. ikisini de dormen tiyatrosundan aparttık
36. oysa işin aslı faslı çok başka
37. efendim dumanların arasından çıkıyorum ki
38. a takımının birer gerçek oyuncusu haline getirdik
39. çok uğraştığı ama bir türlü başaramadığı
40. fırtına gibi esen
41. ama kardeşlik
42. az sonra sarsıcı iddiaların sahibi
43. onu size kısaca bir gösterelim
44. baretta filminden hatırladığımız
45. önde gelen primadonnası
46. a takımı matrak işler de yapar
47. bizim gözlerimize bakarak
48. bizim alkışlarımıza bakarak
49. canlı yayında söyleşeceğiz

50. horoz rolüyle nejat
51. bakın gülbahar ateş
52. şöyle bir görmek ister misiniz
53. gülbahar ateş
54. dumanlar dağalsın
55. bizim şeblemle
56. bu yıldız kıza
57. bu güzel konseri
58. gibi kızımız
59. acımasız
60. heyecanını
61. şunu söylüyorum
62. jipiyle gidiyor
63. sadece
64. bakalım
65. bize reality show izlettireceğiz diye ortalığı salhaneye çevirmeye mezbaya döndürmeye hiç niyetimiz yok

REFERENCES

1. "IAS Report of University of Virginia", 2010, `http://www.web.virginia.edu/iaas/reports/subject/competencies/oralcommun.htm`.

2. "Cancer Facts and Figures - 2010 Report of American Cancer Society", 2010, `http://seer.cancer.gov`.

3. Peterson, G. E., "Parameters of Vowel Quality", *Journal of Speech and Hearing Research*, Vol. 4, pp. 10 – 29, March 1961.

4. Delattre, P., "On the anatomy of intonation", *Lingua*, Vol. 19, pp. 177 – 192, 1963.

5. Diedrich, W. M., "The Mechanism of Esophageal Speech", *Annals of the New York Academy of Sciences*, Vol. 155, No. 1, pp. 303 – 317, 1968.

6. Reich, A., M. McHenry, and F. Minifie, "Acoustical characteristics of intended syllabic stress in excellent esophageal speakers: Linguistic considerations", *Journal of Acoustical Society of America*, Vol. 71, No. S1, p. 56, April 1982.

7. Gandour, J., B. Weinberg, and B. Garzione, "Perception of Lexical Stress in Alaryngeal Speech", *Journal of Speech and Hearing Research*, Vol. 26, pp. 418 – 424, September 1983.

8. Gandour, J. and B. Weinberg, "Perception of Intonational Contrasts in Alaryngeal Speech", *Journal of Speech and Hearing Research*, Vol. 26, No. 1, pp. 142 – 148, 1983.

9. J. Robbins, E. B., H. Fisher and M. Singer, "A comparative acoustic study of normal, oesophageal and tracheoesophageal speech production", *Journal of Speech and Hearing Disorders*, Vol. 49, pp. 202 – 210, 1984.

10. Debruyne, F., P. Delaere, J. Wouters, and P. Uwents, "Acoustic analysis of tracheo-

oesophageal versus oesophageal speech", *The Journal of Laryngology and Otology*, Vol. 108, No. 04, pp. 325 – 328, 1994.

11. Max, L., W. Steurs, and W. Debruyne, "Vocal capacities in oesophageal and tracheoesophageal speakers", *Laryngoscope*, Vol. 106, pp. 93 – 96, 1996.

12. Bellandese, M., J. Lerman, and J. Gilbert, "An Acoustic Analysis of Excellent Female Oesophageal, Tracheoesophageal and Laryngeal Speakers", *Journal of Speech, Language and Hearing Research*, Vol. 44, pp. 1315 – 1320, 2001.

13. Rossum, M. A. v., G. d. Krom, S. G. Nooteboom, and H. Quene, ""Pitch" Accent in Alaryngeal Speech", *Journal of Speech, Language, and Hearing Research*, Vol. 45, No. 6, pp. 1106–1118, 2002.

14. Sharifzadeh, H. R., F. Ahmadi, and I. V. Mcloughlin, "Speech Reconstruction in Post-Laryngectomised Patients by Formant Manipulation and Pitch Profile Generation", *Proc. World Congress on Engineering (WCE) 2009*, Vol. II, London, U.K., July 2009.

15. Ali, R. H. and S. B. Jebara, "Esophageal speech enhancement using source synthesis and formant patterns modification", Damiani, E., K. Yétongnon, P. Schelkens, A. Dipanda, L. Legrand, and R. Chbeir (editors), *Signal Processing for Image Enhancement and Multimedia Processing*, Vol. 31 of *Multimedia Systems and Applications*, pp. 279–288, Springer US, 2008.

16. Aguilar, G., M. Nakano-Miyatake, and H. Perez-Meana, "Alaryngeal Speech Enhancement Using Pattern Recognition Techniques", *IEICE Transactions on Information and Systems*, Vol. E88-D, No. 7, pp. 1618–1622, 2005.

17. Turkmen, H. and M. Karsligil, "Reconstruction of dysphonic speech for synthesizing normally phonated speech", pp. 632–635, apr. 2009.

18. Pozo, A. and S. Young, "Continuous tracheoesophageal speech repair", *Proc. 14th*

European Signal Processing Conference (EUSIPCO '06), Florence, Italy, Sept. 2006.

19. "Lab 9a - Speech Processing (part 1)", 2010, `cnx.org/content/m18086/latest/`.

20. Blue Tree Publishing, Inc., "Vocal Parts", 2000.

21. John R. Deller, J., J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, IEEE Press, 445 Hoes Lane, P.O. Box 1331 Piscataway, NJ 08855-1331, 2000.

22. del Pozo, A., *Voice Source and Duration Modelling for Voice Conversion and Speech Repair*, Ph.D. thesis, Cambridge University Engineering Department, April 2008.

23. Fant, G., "The LF-model revisited. Transformations and frequency domain analysis", *STL-QPSR*, Vol. 36, No. 2-3, pp. 119 – 156, 1995.

24. Doval, B. and C. d´Alessandro, "Spectral correlates of glottal waveform models: an analytic study", *Proc. ICASSP'97*, pp. 446 – 452, Florence, Italy, Sept. 1997.

25. Henrich, N., C. d´Alessandro, and B. Doval, "Spectral correlates of voice open quotient and glottal flow asymmetry: theory, limits and experimental data", *Proc. EUROSPEECH-2001*, pp. 47–50, Florence, Italy.

26. Doval, B., C. C. d´Alessandro, and N. Henrich, "The spectrum of glottal flow models", *Acta Acustica united with Acustica*, Vol. 92, No. 6, pp. 126 – 1046, 2006.

27. George, E. and M. Smith, "Speech Analysis/Synthesis and Modification using an Analysis-by-Synthesis/Overlap-Add Sinusoidal Model", *IEEE Trans. on Speech and Audio Processing*, Vol. 5, No. 5, pp. 389 – 406, 1997.

28. Laroche, J., Y. Stylianou, and E. Moulines, "HNS: Speech Modification Based on a Harmonic plus Noise Model", *Proc. ICASSP'93*, pp. 550 – 553, 1993.

29. Stylianou, Y., J. Laroche, and E. Moulines, "High-Quality Speech Modification Based on a Harmonic plus Noise Model", *Proc. EUROSPEECH'95*, pp. 451 – 454, 1995.

30. L. R. Rabiner, R. W. S., *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, N.J. 07632, 1978.

31. Fujisaki, H. and M. Ljungqvist, "Proposal and evaluation of models for the glottal source waveform", *Proc. ICASSP'86*, pp. 31.2.1 – 31.2.4, 1986.

32. Arslan, L. and D. Talkin, "Voice Conversion by Codebook Mapping of Line Spectral Frequencies and Excitation Spectrum", *Proc. EUROSPEECH'97*, pp. 1347 – 1350, 1997.

33. Türk, O., *New Methods for Voice Conversion*, Master's thesis, Bogazici University, 2003.

34. McLoughlin, I. V., "Line spectral pairs", *Signal Processing*, Vol. 88, No. 3, pp. 448 – 467, 2008.

35. Arslan, L. and D. Talkin, "Line spectrum pair and speech compression", *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 1.10.1 – 1.10.4, San Diego, Calif., 1984.

36. "The American Family Physician Web Site", 2010, `http://www.aafp.org/afp/980600ap`.

37. "The Internet Encyclopedia of Science", 2010, `http://www.daviddarling.info/encyclopedia/L/larynx.html`.

38. "University of Pittsburgh Voice Center", 2010, `http://www.pitt.edu/~crosen/voice/anatomy2.html`.

39. "Larynx - Wikipedia, the free encyclopedia", 2010, `http://en.wikipedia.org/`

wiki/Larynx.

40. "Laryngeal Stenosis: eMedicine Otolaryngology and Facial Plastic Surgery", 2010, http://emedicine.medscape.com/article/867177-overview.

41. "Laryngectomy Basics", 2010, http://www.larynxlink.com/Library/faqs/FAQ16.htm.

42. "Laryngeal and Hypopharyngeal Cancer - American Cancer Society", 2010, http://en.wikipedia.org/wiki/Larynx.

43. "Cancer Network Web Site", 2010, http://imaging.ubmmedica.com/cancernetwork/journals/oncology/images/o0006ef2.jpg.

44. "Blom-Singer Voice Prostheses - InHealth Technologies", 2010, http://www.inhealth.com/featuredprdvppage1new.htm.

45. Singer, M. I. and E. D. Blom, "Tracheoesophageal puncture: A surgical prosthetic method for post laryngectomy speech restoration", 1979, third International Symposium Plastic Reconstructive Surgery of the Head and Neck, New Orleans.

46. "Cancer Network Web Site", 2010, http://imaging.ubmmedica.com/cancernetwork/journals/oncology/images/o0006ef3.jpg.

47. "Cancer Network Web Site", 2010, http://imaging.ubmmedica.com/cancernetwork/journals/oncology/images/o0006ef1.jpg.

48. Türk, O., *Cross-Lingual Voice Conversion*, Ph.D. thesis, Bogazici University, 2007.