University of Deusto, Bilbao, Spain

Doctoral Thesis

Enhancement of Esophageal Speech using Signal Processing algorithms on Source Signal and Vocal Tract Filter

Author:
Rizwan Ishaq

Supervisor:
Dr. Begoña García Zapirain

A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy

in the

University of Deusto
for Doctoral Program in Computer and Telecommunication Engineering

November 2015

Declaration of Authorship

I, Rizwan Ishaq, declare that this thesis titled, 'Enhancement of Esophageal Speech using Signal Processing algorithms on Source Signal and Vocal Tract Filter' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

"Learn to Fail, Fail to learn."

Abstract

The speech an essential component for daily life communication sometimes alter due to laryngeal cancer treatment. The advanced stage treatment for laryngeal cancer is total laryngectomy. The one of the consequences of total laryngectomy is that normal speech production destroyed and alternative speech production are needed. The Esophageal Speech (ES) is one of the alternative speech production method after total laryngectomy. The ES uses esophagus as an alternative to larynx, and the air source comes from mouth to the lower esophagus, and then release back which vibrates the esophagus and provides voicing source to the vocal tract filter. The produced speech by this method has low quality and low intelligibility due to irregular voicing source and altered vocal tract filter. This thesis, therefore presents an enhancement method for ES by transforming the source and vocal tract filter components into normal speech components. The system in the thesis, first decompose the ES into source and vocal tract filter components using Iterative Adaptive Inverse Filtering (IAIF), and then transforms these components into normal speech components. The source most effected, is first decomposed into fundamental frequency F_0 curve, Harmonic to Noise Ratio (HNR) and source spectrum components. The natural glottal pulse computed from nomral speech is used with normal speech F_0 curve and HNR along with original source spectrum for transformed source signal. The vocal tract filter is transformed by smoothing the vocal tract spectral peaks, and then shifting these spectral peaks to lower frequencies using second order Frequency Warping Function (FWF). The spectral peaks widths are then enlarge to make it more closure to natural speech. The system is evaluated using subjective listening tests and objectively using HNR on the Spanish ES vowels /a/, /e/, /i/, /o/, /u/, and 28 mostly used Spanish words. The subject listening tests using MOS and preference score have shown that proposed system MOS always between 3 to 4, and the preference for all the processed sample is more than 50%. The objective result using HNR has shown 10 to 15 dB improvement over the original ES samples.

Acknowledgements

I am thankful to my advisor, Dr. Begoña García Zapirain, for her guidance, support, encouragement and her devotion of time in this research. Special thanks to my colleagues Gonzalo Equíluz Pérez, Fernando Jorge, Álvaro Javier Muro, and all from the duestotech-evida lab who supported me throughout my research period. I highly acknowledge all those who participated in the extensive listening tests. I am highly grateful to Dr. Paavo Alku, Dr. Dhananjaya Gowda, and Tuomo Raitio for providing me helpful ideas and discussion during my stay in Finland. I am highly thankful to my parents for their support throughout my life.

Contents

List of Figures

List of Tables

Abbreviations

AR	autoreggressive
CC	complex cepstrum
CP	closed phase
CIQ	closing quotient
CQ	closed quotient
DAP	discrete all-pole modeling
DSP	digital signal processing
DYPSA	dynamic programming projected phase-slope algorithm
EGG	electroglottography
EL	electrolarynx
ES	esophageal speech
GCI	glottal closure instant
FIR	finite impulse response
GIF	glottal inverse filtering
GOI	glottal opening instant
HNR	harmonic to noise ratio
HRF	harmonic richness factor
IAIF	iterative adaptive inverse filtering
IIR	infinite impulse response
LF	Liljencrants-Fant
LP	linear prediction
LPC	linear prediction coding
LTI	linear time-invariant
MOS	mean opinion score
STE	short-term energy

TES	tracheo-esophageal speech
WLP	weighted linear prediction
ZZT	zeros of the z transform

Dedicated to my beloved parents

Chapter 1

Introduction

The speech is our daily source of communication, produced by modulated air source and shaped by oral cavity. The air source comes from the lungs to the vocal folds which are resided in larynx. The lungs and larynx are connected through trachea. The air source forces the vocal folds to open and close depending upon the type of source. The vocal folds are held open for the unvoiced source, while for the voiced source, it opens and closes periodically. This voiced/unvoiced source is then spectrally shaped by oral cavity consists of pharynx, vocal tract, and nasal cavity. The spectrally shaped voicing source is then radiated into air by lips. In short, the essential parts for speech production are air source, voicing source and oral cavity (vocal tract). The larynx produces the voicing source, and considered an essential component for the speech production.

The laryngeal cancer uncommon type of cancer increased in the last few years due to excessive use of tobacco and alcohol (although no study available among the correlation of tobacco and alcohol use for laryngeal cancer). According to American Society [1], there are 12,360 cases reported in USA, and 28,000 in European Union (EU) in 2012 [2]. There are treatments to laryngeal cancer, such as, chemotherapy, radiotherapy, partial and total laryngectomy. The chemotherapy and radiotherapy are the mostly used treatments in the modern age, but still the advanced stage laryngeal cancer can not be treated with these treatments. The partial laryngectomy is sometime can help to reduces the laryngeal cancer, but the last stage treatment of laryngeal cancer still needs total laryngectomy to save the life of the patient. The larynx of the patient removed in the total laryngectomy. The air pathway from lungs to the mouth no more available, and for the breathing purpose, a hole on neck called stoma, created and trachea one end connected to the stoma.The consequences of the total laryngectomy are extreme, such as, breathing pathway altered, food intake effected, and the most severe is the speech production. The larygectomee (patient who went through total laryngectomy) can not produced the

speech, as the speech production essential components such as air source, and larynx are not available. However to regain the speech production ability, the alternative methods are used, which use different body organs as an alternative to normal speech production organs. The alternative of larynx for the larygectomy is created using the esophagus or external devices. There are three mostly used method for speech regaining after total laryngectomy, i) Esophageal Speech (ES), Tracheo-Esophageal Speech (TES), and Electrolarynx (EL). The ES and TES both use the Pharyngo-Esophagus Segement (PES) in esophagus as voicing source or alternative to larynx. Although ES and TES both use the same voicing source, the air source is still different for each other. The air source for the TES comes from the lungs by diverting the air from trachea to esophagus by inserting the one way volve between trachea and esophagus. The one-way volve prevents the food to enter into trachea. The air source however comes from the mouth by inhaling air into the esophagus, and then exhaling the air which vibrates the PES segment for the voicing source. The EL is the most simplest method and does not use any air source for speech production, instead it uses external vibrating devices for voicing source.

1.1 Justification

The different speech production method have different advantages and disadvantages, such as, the speech production using EL is the most easiest way of producing speech for laryngectomee, but it has disadvantage of being more robotic (machine like sound), and the use of external devices. The TES is the more closer to the normal speech, but it needs complex surgery for inserting volve between trachea and esophagus and need cleaning of that volves daily. The ES is the most natural way of producing speech after total laryngectomy, despite it takes time to learn, as well the low air pressure produces the low quality speech. The ES, therefore is mostly used speech production method after total laryngectomy, because it does not require surgery as in TES, and external devices as in EL. Despite the preferred method, the ES has the following deficiencies in comparison to natural speech, i) the air source from the mouth has low air pressure ii) the irregular shape of PES and low air source pressure generates irregular voicing source, iii) the vocal tract also effected due to total laryngectomy, i.e. shortening of vocal tract. Due to these deficiencies, the produced speech has low intelligibility and quality. The speech almost sounds like a burp. In a sense of speech components, the ES has low fundamental frequency, almost no harmonics in voicing source and corresponds whisper speech voicing source, and vocal tract spectral peaks are moved higher in the frequency. The produced speech needs special algorithms for enhancing its intelligibility, which transform the voicing source and vocal tract filter into natural speech components. In order to address these problems of ES, special signal processing algorithms are needed

which transform the ES speech in to almost normal speech. This thesis therefor uses the signal processing algorithms for decomposing the ES into source and filter components and then transforms these components into normal speech components for better and intelligible speech, which make the life of laryngectomee easy. The speech signal processing industry have not put effort on this type of speech signals, so this thesis hence provided the algorithms to speech signal processing industry to make the speech coders for this type of speech signals for the betterment of the life of the laryngectomee.

1.2 Hypothesis

It is possible to enhance the intelligibility and quality of low quality ES using the signal processing algorithms, which transforms the ES source and filter components to the normal speech source and filter components using the natural glottal pulse and Frequency Warping Function (FWF).

To deal with transformation of ES source and filter components into natural speech components, this thesis uses source-filter theory of speech production [3], which describes the speech as the combination of source and filter. The thesis assumes that ES can be faithfully decomposed into its source and filter components according to [3]. After decomposition these components are processed independently. The source components are transformed into natural speech components by borrowing natural glottal pulse, while the filter components are transformed into natural filter components using the frequency warping function.

1.3 Objectives

The overall objective of this thesis is to enhance the ES quality and intelligibility by transforming the ES source and filter components into normal speech source filter components. The following are the main objectives addressed and solved in this thesis;

- Build a high quality pathology speech database for experimentation
- Design a method for decomposing ES into source and vocal tract components.
- Design a new algorithm for transforming the ES source signal into normal speech source signal.

- Design a novel algorithm for transforming the ES vocal tract filter into normal speech vocal tract filter.
- Evaluate the quality of enhanced ES subjectively and objectively.

The first step in enhancing the ES is to decompose the ES into its source and filter components, which is done by automatic inverse filtering method Iterative Adaptive Inverse Filtering (IAIF) [4]. After analyzing the source and filter components, the problems of ES source and filter are solved by transforming them into normal speech source-filter components. For this purpose, separate algorithms are designed for source and filter. The source is transformed using the natural glottal pulse extracted from the normal speech, while the filter is transformed using the Frequency Warping Function (FWF), which solves the deficiencies of ES filter. At the end the transformed source and filter components are synthesized for enhanced better quality ES. The proposed algorithms are assessed using the subjecting listening tests, and objectively using Harmonic to Noise Ratio (HNR). The spectrogram is also used to assess the system visually.

1.4 Structure of Thesis

The thesis structure is as follows;

- Chapter 1 provided the introduction of the thesis.
- Chapter 2 presented the background study and state of the art review related to the ES in detailed.
- Chapter 3 then provided the proposed system to address the problem of ES by decomposing the ES into source and filter components and then transforming these components into normal speech components. The source signal is modified using the natural glottal pulse, and the vocal tract filter is transformed using the Frequency Warping Function (FWF).
- Chapter 4 evaluated the proposed system, and presented the results.
- Chapter 5 then concluded the thesis and presented the future lines for the proposed system.

Chapter 2

State of the Art Review

Before going into detail of ES speech production method, it is necessary to first explain the natural speech production mechanism, and then describe the ES in comparison to natural speech. This chapter first describes the the natural speech production mechanism, and then in comparison to natural speech production ES production methods is introduced.

2.1 Speech Production Methods

2.1.1 Natural Speech

The natural speech production mechanism can be divided into following essential components (Figure 2.1):

- Lungs
- Larynx
- vocal tract

The primary purpose of lungs is to provide the air source to the larynx for voicing source. The larynx and lungs are connected through trachea, and air passes through trachea to larynx. The larynx has the vocal folds. The air source pressure from lungs forces the glottis in vocal folds to open and close depending on the type of phonation. In voiced phonation, air source, modulated by opening and closing of glottis, provides periodic source signal, while for unvoiced phonation, glottis held open, and the source signal corresponds to noise signal. The source signal then passed through the vocal tract.

The vocal tract consists of pharyngeal, oral, and nasal cavities and shaped the source signal using spectral resonance and anti-resonance (Figure 2.1). Finally, the spectrally shaped source (glottal) signal radiated into air through lips or nose. For the convinces, the modulated air source, through larynx, is called, glottal source, or voicing source, and the combination of oral, pharynx and nasal cavities considered as the vocal tract. The glottal source, a main component for the phonation, provides the voiced source, i.e. quasi-periodic vibration of glottis, and unvoiced source, when the glottal source is noisy.

In comparison to natural speech production, the speech after total laryngectomy, uses different organs for production. The Figure 2.2 shows the difference between laryngectomee and non-laryngectomee organs. The larynx is missing, and the air source from lungs do not pass through the vocal tract, instead larygectomee breath through the hole on the neck called stoma. The equivalence of vocal folds, and air source are needed for speech restoration. The esophagus and external devices are used for this purpose. Two methods, Esophageal Speech (ES), and Tracheo-Esophageal use esophagus for voicing source generation, with a different air source. Electrolayrnx (EL) uses the external devices, without any air source for voicing source.

2.1.2 Electorlarynx (EL)

The EL uses the external vibrating device for voicing source for speech production. The EL does not use any air source for voicing source. Figure 2.3 shows the EL speech production process. The external vibration device is placed against the neck, and vibration of device provides the voicing source. EL is a easiest method for speech restoration after total laryngectomy, but it sounds robotic.

2.1.3 Treacheo-Esophageal Speech (TES)

The another type of speech production method after total laryngectomy which uses the esophagus as shown in Figure 2.4. The Paryngeo-esophagus (PE) segment in esophagus is used as a voicing source generator. The air source from lungs is diverted to the esophagus using the one way valve inserted between trachea and esophagus by surgery. The vibration of PE segment provides the voicing source to the vocal tract for speech signal. TES is more closer to the natural speech production, because the air source pressure from lungs is higher. The problem with TES is that it requires surgery, and continues cleaning of valve, inserted for air diverting to esophagus.

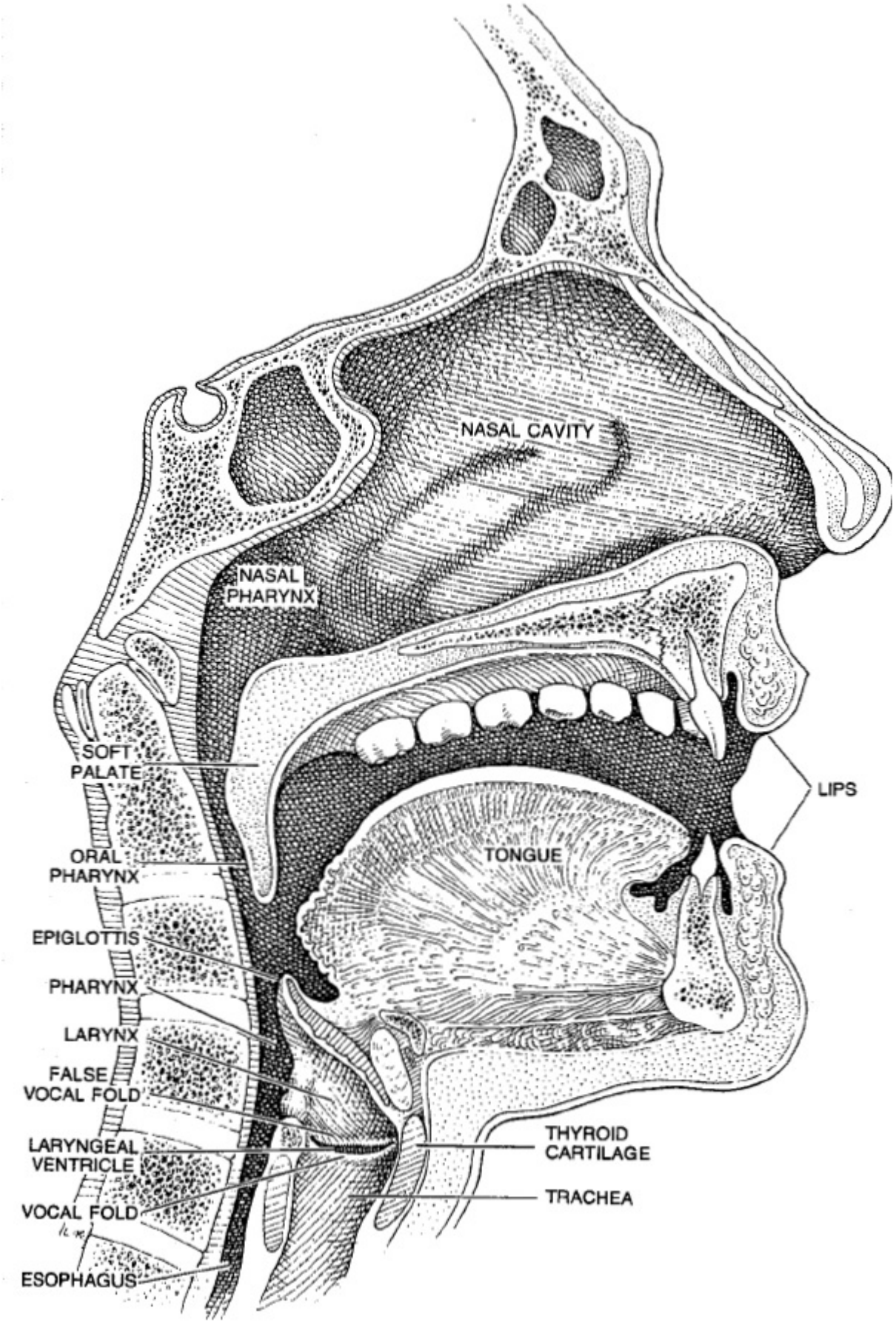

FIGURE 2.1: Speech production organs (adapted from [5])

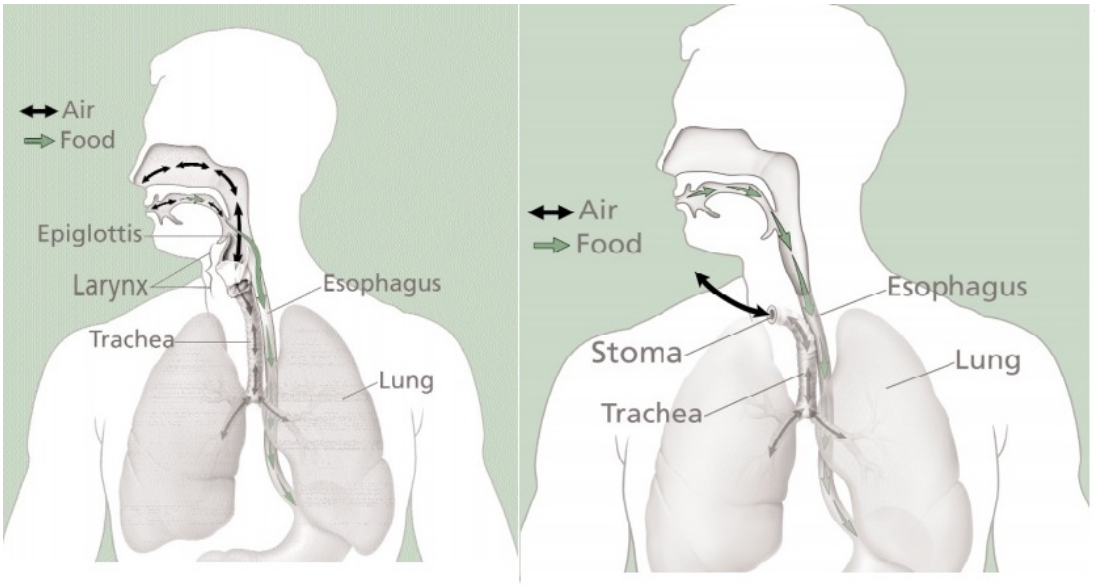

FIGURE 2.2: Organs before and after total laryngectomy (adapted from [6])

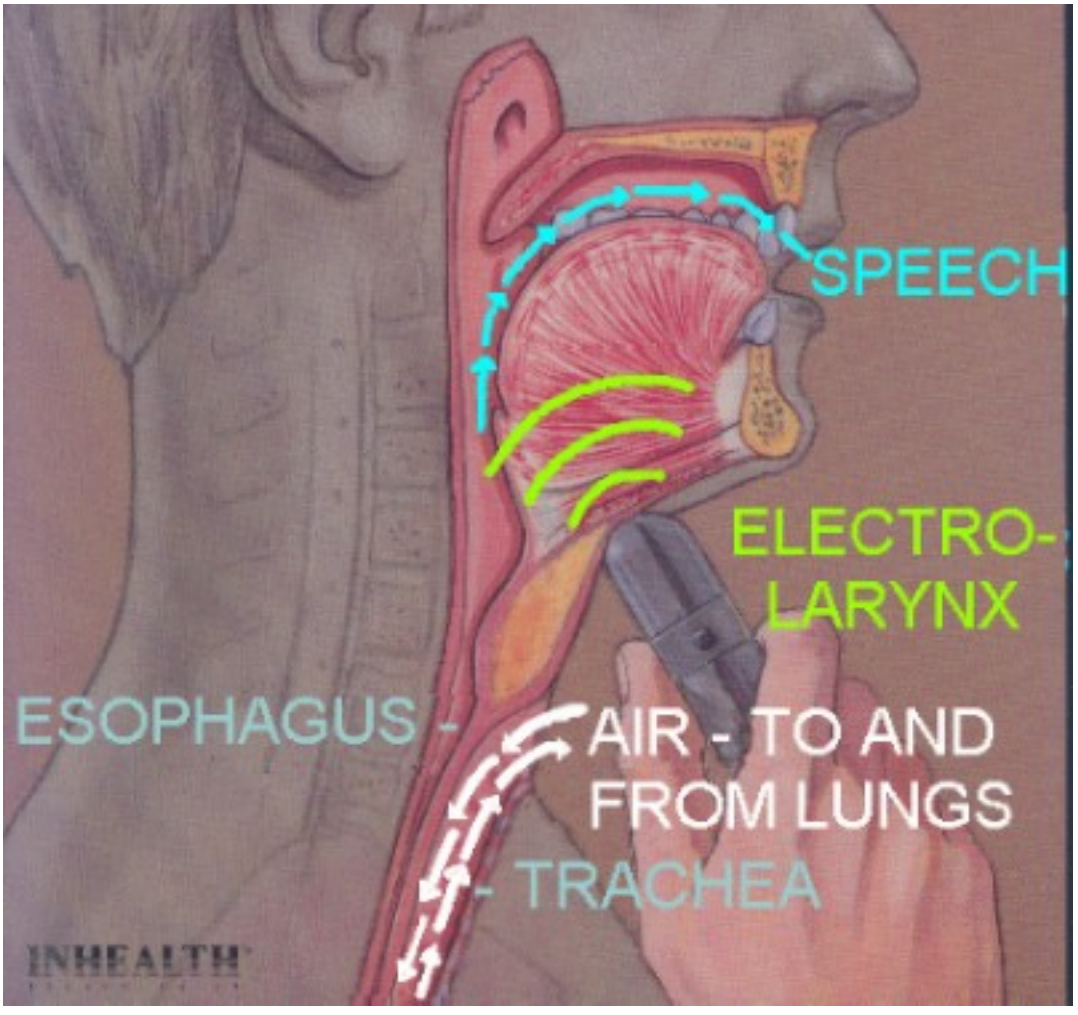

FIGURE 2.3: Electrolaryx speech production (adapted from [7])

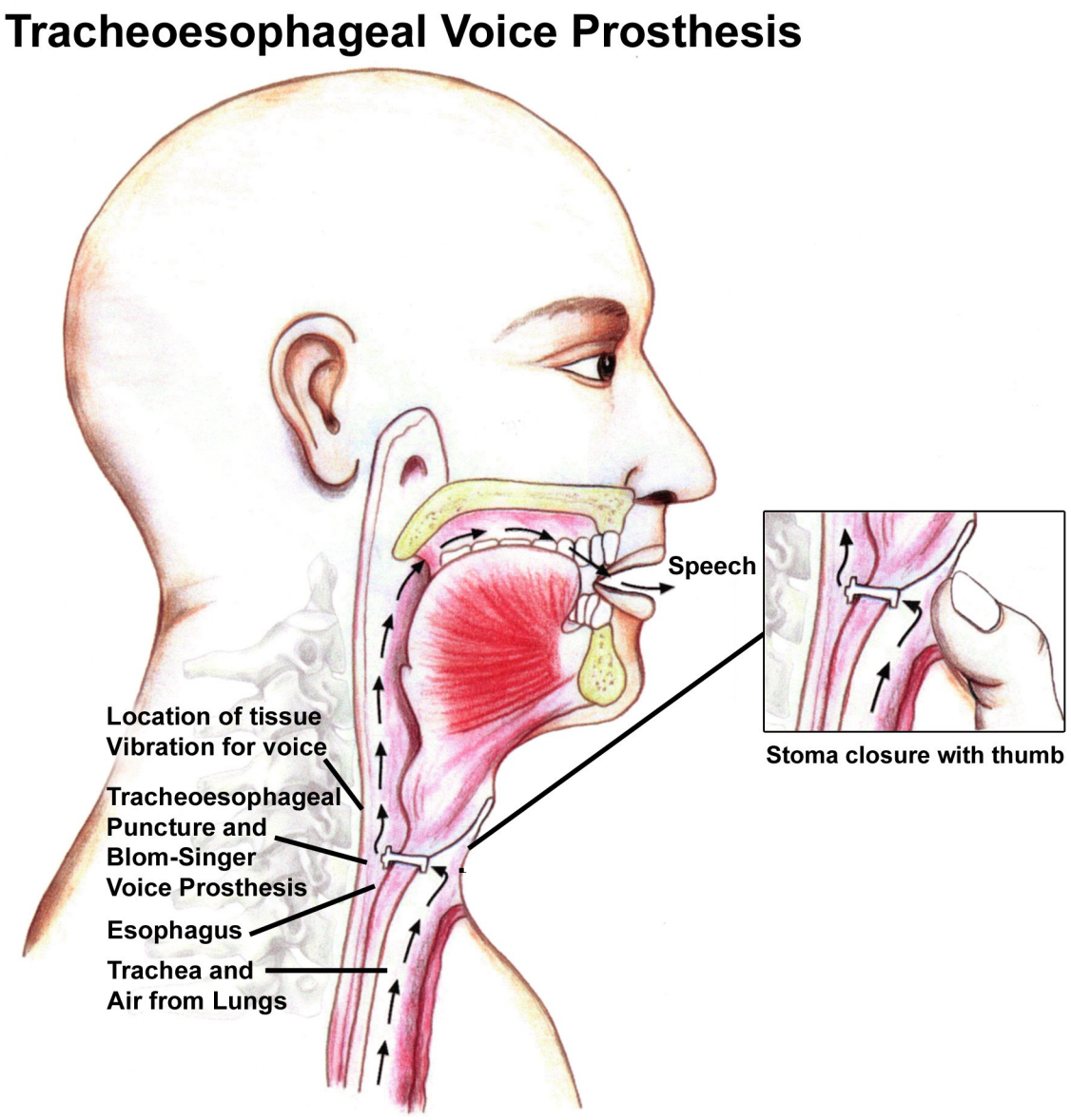

FIGURE 2.4: Treacheo-Esophageal Speech production (TES) (adapted from [7])

2.1.4 Esophageal Speech (ES)

Esophageal Speech (ES) is another speech restoration method used after total laryngectomy. The speech production process for ES is similar to TES, the difference is only the air source. To avoid the surgery, ES uses a air source by inhaling air through mouth to the lower part of esophagus, and then exhale back, which vibrates the PE segment, and provides a voicing source for speech production as shown in Figure 2.5. Although ES air source pressure is low as compared to TES, but it is preferred method because it does not need any surgery or external devices for source generation [8].

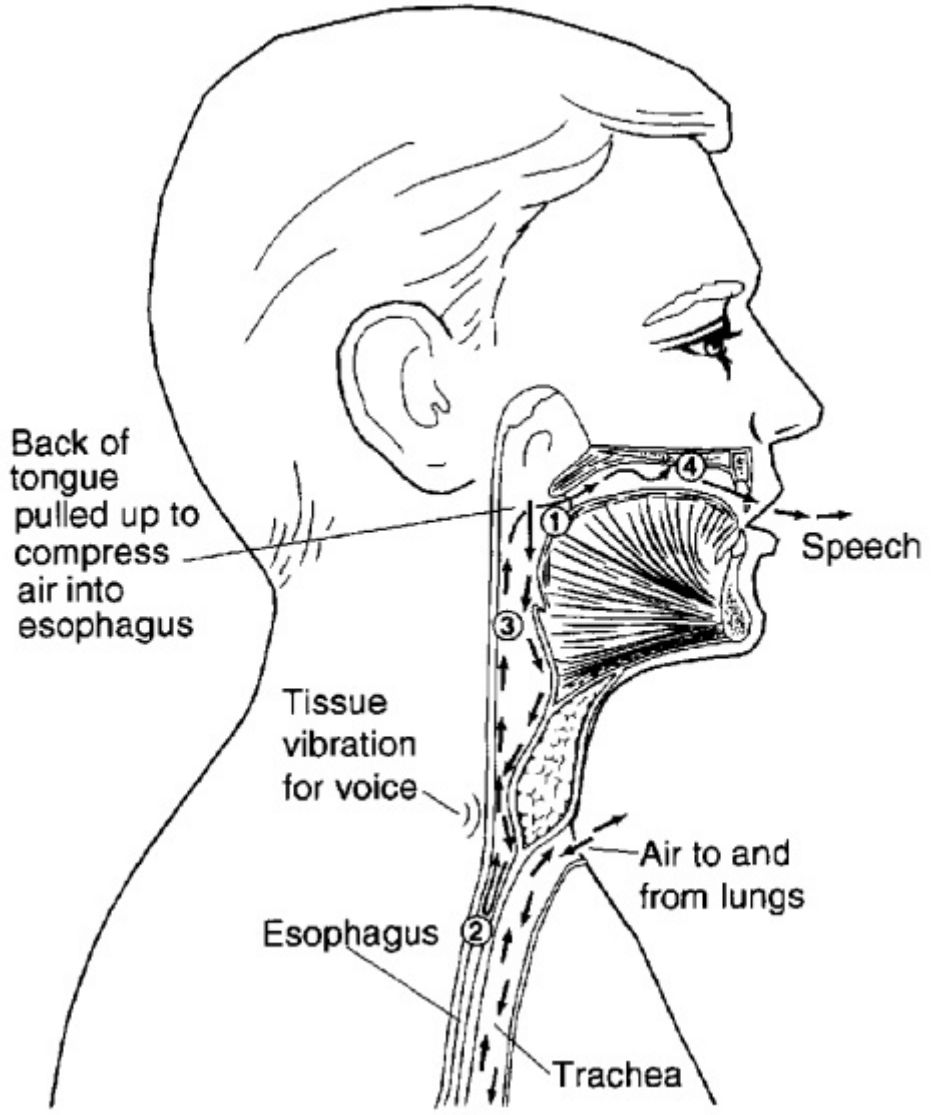

FIGURE 2.5: Esophageal Speech (adapted from [8])

2.2 Speech Modeling

As this thesis only deal with ES, therefore ES will be discussed in a sense of natural speech production. The air source for normal speech comes from lungs, but for ES it comes from mouth by inhaling air. The neo-glottal in PE segment, vibration provides voicing source, while in normal speech glottis in vocal folds are used for this purpose. The vocal tract is considered similar to that of normal speech. So the ES can be modeled as the linear source-filter model of speech production [3]. The speech production mechanism is a complex and non-linear process, but the speech can be modeled using the linear source filter model for the short length frames of speech (i.e. 30-ms) based on that vocal tract and source of speech are independent and decomposable. Therefore, based on simple assumption of source and vocal tract separability, short segment of speech

is considered as the output of Linear Time Invariant (LTI) system, whose input is the source signal [3] (Figure 2.6). Based on this LTI system of speech production [3] shown

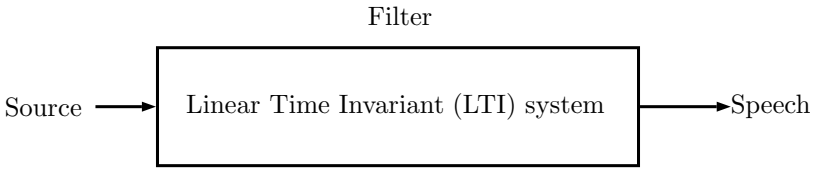

FIGURE 2.6: Linear Time Invariant (LTI) system

in Figure 2.7, the speech signal $s[n]$ is convolution of source signal, vocal tract, and lip radiation.

$$s[n] = g[n] * v[n] * r[n] \tag{2.1}$$

where $g[n]$, $v[n]$, $r[n]$ and $*$ are source signal, vocal tract, lip radiation and convolution operator, respectively. In z-domain;

$$S(z) = G(z)V(z)R(z) \tag{2.2}$$

where $G(z)$, $V(z)$, $R(z)$, and $S(z)$ are the transfer functions of $g[n]$, $v[n]$, $r[n]$ and $s[n]$, respectively. The source signal for voiced speech in ES is consists of periodic impulses

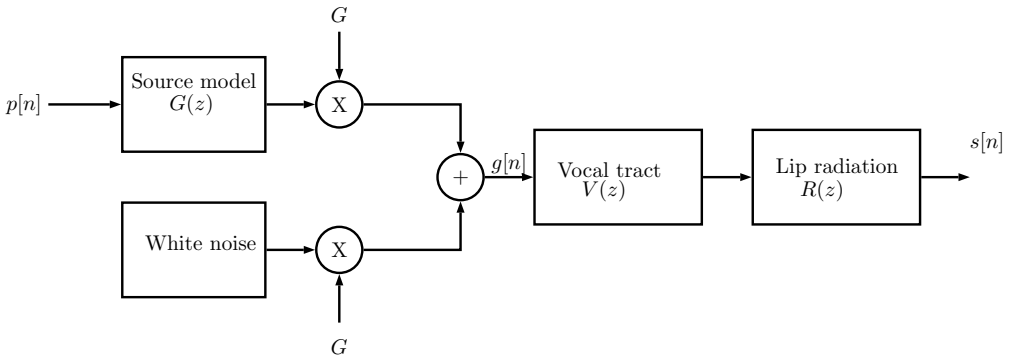

FIGURE 2.7: The source-filter model of speech for voiced and unvoiced speech

$p[n]$ provided by vibration of Pharyngo-esophagus (PE) segment, and filtered by the source model, and for unvoiced speech it is white Gaussian noise. The vocal tract is an all-pole model of voiced speech, and the lip radiation is differential filter.

2.2.1 Speech components

2.2.1.1 Lip radiation

The microphones measure the speech as pressure waves, and most of them operate in the far filed, and the lip radiation influenced the speech signal. Therefore the lip radiation is also present in the speech signal and acoustically it is approximated with first-order difference filter;

$$R(z) = 1 - \alpha z^{-1}, \quad 0.96 < \alpha < 1 \tag{2.3}$$

where α is a radiation constant. The frequency response and corresponding pole of the filter is shown in Figure 2.8.

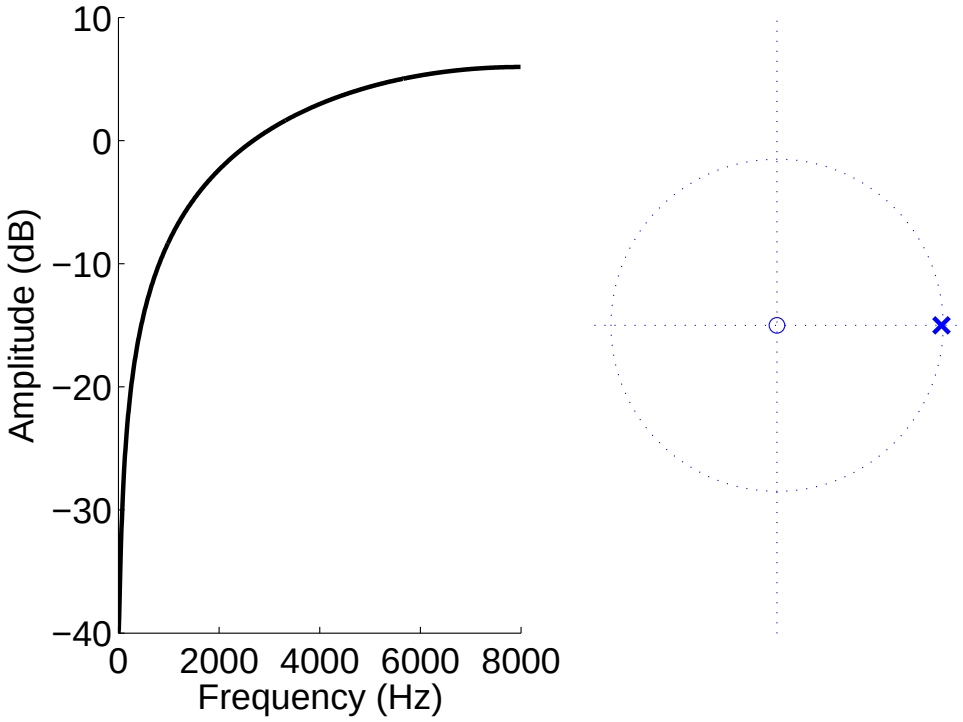

FIGURE 2.8: Lip radiation frequency response and its correspond pole for $\alpha = 0.99$

2.2.1.2 Vocal tract

The vocal tract an essential component for speech production, can be divided into three main parts i) pharynx, ii) oral cavity, and iii) nasal cavity. The Figure 2.9 shows the vocal tract in term of cylindrical tube of varying cross-section [9, 10]. Typically, the vocal tract has the length of around 17 cm for adult male, and 14 cm for female. Acoustically, the vocal tract is modeled by formants (poles) and anti-formants (zeros).

For the simplicity, most of the voiced sound can be modeled by all-pole model, only some nasal sound introduced zeros in the sound. Hence for simplicity, it can be modeled as an all-pole model;

$$V(z) = \frac{1}{1 + \sum_{k=1}^{P} a_k z^{-k}} \tag{2.4}$$

where a_k are the all-pole models prediction coefficients of order P for vocal tract $V(z)$. Following assumptions are needed for all-pole modeling of vocal tract to be valid [11];

- tube cross-sectional area is considered constant
- there is no air turbulent within tube
- glottis and vocal tract are linearly separable

The linear prediction coefficients of vocal tract $V(z)$ can be approximated by different methods. Some of them are given below.

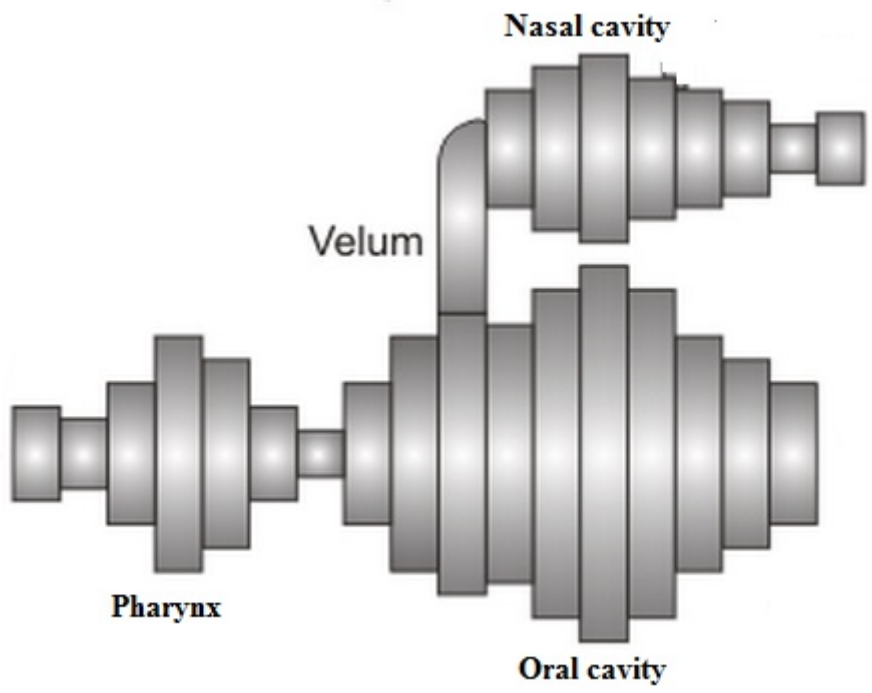

FIGURE 2.9: Vocal tract (adapted from [12])

2.2.1.2.1 Linear Prediction Coding (LPC)

The Linear Prediction of speech signal is commonly used method for speech signal processing, used for estimating speech parameters, such as, pitch, formant frequencies, vocal tract filter etc. The method initially developed for speech coding, but afterward it

has been used extensively, for speech analysis, recognition, enhancement etc. The basic idea of linear prediction is that, a current speech sample is linear combination of past speech samples [13, 14]. According to linear prediction,

$$\hat{s}[n] \approx \sum_{k=1}^{P} a_k s[n-k] \tag{2.5}$$

where a_k are the linear prediction coefficients of order P. In order to find the prediction coefficients, the error between $s[n]$ and estimated signal $\hat{s}[n]$ is calculated as;

$$e[n] = s[n] - \hat{s}[n] = s[n] - \sum_{k=1}^{P} a_k s[n-k] \tag{2.6}$$

The prediction coefficients are estimated by minimizing the mean square error over a short segment of speech signal;

$$\epsilon = \sum_{n=1}^{N} e[n]^2 \tag{2.7}$$

$$\epsilon = \sum_{n=1}^{N} (s[n] - \hat{s}[n])^2 \tag{2.8}$$

$$\epsilon = \sum_{n=1}^{N} [s[n] - \sum_{k=1}^{P} a_k s[n-k]]^2 \tag{2.9}$$

where N is the segment length. The error is minimized by setting ϵ derivative to zero with respect to a_k;

$$\frac{\partial \epsilon}{\partial a_i} = 0, \quad i = 1, 2, 3, \ldots, P \tag{2.10}$$

$$\sum_{n=1}^{N} 2\{s[n] - \sum_{k=1}^{P} a_k s[n-k]\}[-s[n-i]] = 0 \tag{2.11}$$

$$\sum_{n=1}^{N} s[n]s[n-i] - \sum_{n=1}^{N}\sum_{k=1}^{P} a_k s[n-k]s[n-i] = 0 \tag{2.12}$$

$$\sum_{k=1}^{P} a_k \sum_{n=1}^{N} s[n-k]s[n-i] = \sum_{n=1}^{N} s[n]s[n-i], \quad \text{for } i = 1, 2, 3, \ldots, P \tag{2.13}$$

The autocorrelation function is given as;

$$\phi[k] = \sum_{n=1}^{N} s[n]s[n+k] \tag{2.14}$$

Inserting autocorrelation function, the equation become;

$$\sum_{k=1}^{P} a_k \phi[k-i] = \phi[i], \quad \text{for } i = 1, 2, 3, \ldots, P \tag{2.15}$$

$$a_1\phi[0] + a_2\phi[1] + a_3\phi[2] + \ldots + a_P\phi[P-1] = \phi[1] \tag{2.16}$$

$$a_1\phi[1] + a_2\phi[0] + a_3\phi[1] + \ldots + a_P\phi[P-2] = \phi[2] \tag{2.17}$$

$$a_1\phi[2] + a_2\phi[1] + a_3\phi[0] + \ldots + a_P\phi[P-3] = \phi[3] \tag{2.18}$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots = \ldots \tag{2.19}$$

$$a_1\phi[P] + a_2\phi[P-1] + a_3\phi[P-2] + \ldots + a_P\phi[0] = \phi[P] \tag{2.20}$$

In matrix form;

$$\begin{bmatrix} \phi[0] & \phi[1] & \ldots & \phi[P-2] & \phi[P-1] \\ \phi[1] & \phi[0] & \ldots & \phi[P-1] & \phi[P-2] \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \phi[P-1] & \phi[P-2] & \ldots & \phi[1] & \phi[0] \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_P \end{bmatrix} = \begin{bmatrix} \phi[1] \\ \phi[2] \\ \vdots \\ \phi[P] \end{bmatrix} \tag{2.21}$$

in simplified form it is;

$$\phi\vec{a} = \vec{\phi_0} \tag{2.22}$$

The linear prediction coefficients are then given as;

$$\vec{a} = \phi^{-1}\vec{\phi_0}, \quad \vec{a} = [a_1, a_2, \ldots, a_P] \tag{2.23}$$

where ϕ and $\vec{\phi}$ are autocorrelation matrix and vector, respectively.

2.2.1.2.2 Weight Linear Prediction (WLP)

Another method to estimate the vocal tract coefficients [15], is similar to linear prediction, but with a weighting function, to focus on the close phase estimation. The only different between linear prediction and weighted linear prediction is the energy function;

$$\varepsilon = \sum_{n=1}^{N} [s[n] - \sum_{k=1}^{P} s[n-k]]^2 W[n] \tag{2.24}$$

where $W[n]$ is a weighted short time energy function and is given as [15, 16];

$$W[n] = \sum_{k=1}^{M} s^2[n-k] \tag{2.25}$$

where M is the number of samples for energy estimation. Solving ε give the desired prediction coefficients.

2.2.1.2.3 Stabilized Weighted Linear Prediction (SWLP)

The WLP which compute the all-pole model of speech using weighting window [17], and is improved version of linear prediction all-pole model. Although the model produces better result, but it does not guarantee the stability of all-pole model. To introduce the stability in WLP, Stabilized Weighted Linear Prediction (SWLP) [15] is used which modifies the weighting function so that the all-pole model guarantees stability.

2.2.1.2.4 Extended Weighted Linear Prediction (XLP)

The eXtended Weighed Linear Prediction (XLP) is another modification to WLP, where two-dimensional weighting function is used to handle to problems of instability of all-pole model present in WLP and LP [18]. The two-dimensional weighting function $Z[n, j]$ for XLP is

$$Z[n,j] = \frac{m-1}{m} Z[n-1,j] + \frac{1}{m}(|s[n]| + |s[n-j]|), \quad \text{where } m = 20 \tag{2.26}$$

Using the $Z[n, j]$,

$$\sum_{k=1}^{P} a_k \sum_{n=1}^{N} Z[n,k]s[n-k]Z[n,j]s[n-j] = \sum_{n=1}^{N} Z[n,0]s[n]Z[n,j]s[n-j], \quad 1 \leq j \leq p \tag{2.27}$$

The solution to above minimization of energy equation gives a_k.

2.2.1.2.5 Discrete all-pole (DAP) model

The Discrete all-pole (DAP) is another method to estimate the vocal tract filter accurately, which uses the Itakura-Saito (I-S) error measure instead of mean square error in linear prediction modeling by matching the appropriate autocorrelation function [19].

2.2.1.3 Source signal

The source signal which is produced by the neoglottis in ES and by glottis in normal speech produces different types of voicing source, depending on type of phonation. For unvoiced phonation it is white Gaussian noise and for the voiced speech, the source

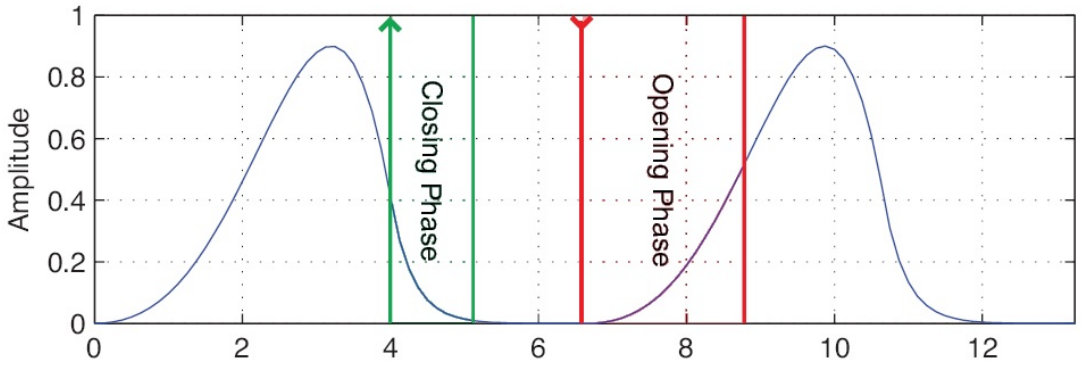

FIGURE 2.10: Glottal flow (adapted from [11])

signal is consists of quasi-periodic cycles, and each cycle consist of opening and closing phase as shown in Figure 2.10.

This one glottal cycle can be model by different source model, and simplest one is the two-pole model [20, 21]:

$$G(z) = \frac{1}{1 + \sum_{k=1}^{2} \alpha_k z^{-k}}, \quad \alpha_1 \text{and } \alpha_2 \approx 1 \tag{2.28}$$

where α_k are the two real poles near to 1. According to the Figure 2.11, the glottal flow and its derivative can be modeled using following timing instants [22];

- t_s start of glottal pulse
- t_i time at maximum of the glottal pulse derivative
- t_p time at maximum of glottal pulse
- t_e time at minimum of glottal pulse derivative
- t_a return phase duration
- t_c glottal pulse closing time
- T_0 pitch period of pulse

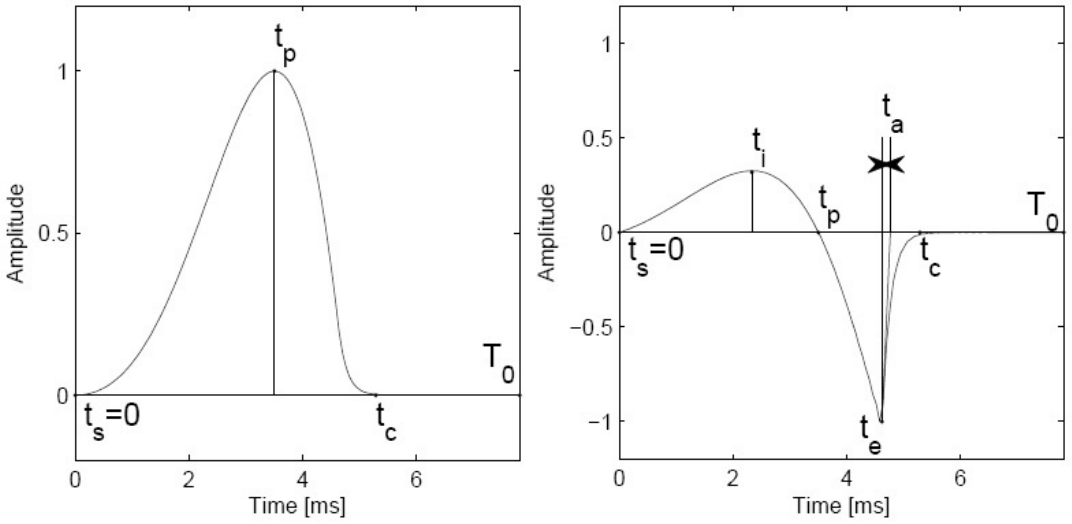

FIGURE 2.11: Glottal flow and its derivative timing instants

2.2.1.3.1 Rosenberg source model

Based on these timing instant, Rosenberg model [22, 23] model approximates the glottal source wave form according to following mathematical relation:

$$g(t) = \begin{cases} t^2(t_e - t) & \text{if } 0 < t < t_e = tc \\ 0 & \text{if } t_c < t < T_0 \end{cases} \tag{2.29}$$

2.2.1.3.2 Fant source model

Another model based on two sinusoidal curves and called Fant model [22, 24] is given accordingly:

$$g(t) = \begin{cases} \frac{1}{2}(1 - cos(\omega_g t) & \text{if } 0 < t < t_p \\ K.cos(\omega_g(t - t_p)) - K + 1 & \text{if } t_p < t < t_c \\ 0 & \text{if } t_c < t < t_0 \end{cases} \tag{2.30}$$

where $\omega_g = \frac{\pi}{t_p}$, and K shape controlling parameter.

2.2.1.3.3 Liljencrants-Fant source model

One of the mostly used source model for speech signal processing is the Liljencrants-Fant model [25], and model using the following mathematical relation:

$$g'(t) = \begin{cases} -E_e e^{a(t-t_e)} \frac{sin(\frac{\pi t}{t_p})}{sin(\frac{\pi t_e}{t_P})} & 0 \leq t \leq t_e \\ -\frac{E_e}{\epsilon t_c}(e^{-\epsilon(t-t_e)} - e^{-\epsilon(t_0-t_e)}) & t_e \leq t \leq t_0 \end{cases} \tag{2.31}$$

The parameters α and ϵ are estimated by solving the above equation as;

$$\int_0^{t_0} g'(t)dt = 0 \tag{2.32}$$

Some other methods are also available, such as Transform-LF [26] used to model the LF model using the R parameters. The Causal-Anticausal Linear Model (CALM) [27], modeled the source signal in spectral domain using two filters, one to modeled the open phase using anti-causal poles pair, and return phase by causal real pole. The Klatt [28] and Fujisaki [29] are also, but not so common.

2.2.2 Source and vocal tract decomposition

The glottal source or simply source and vocal tract decomposition is advantageous for many speech processing application, such as, speech coding, speech synthesis, speech recognition, etc. Both components provide different characteristics of speech signal. In order to capture, the variation, it is compulsory to process the source and vocal tract independently of each other.

2.2.2.1 Linear prediction source-filter decomposition

The linear prediction coding/analysis is the most used method in speech processing [14], where the speech current samples is estimated from the past P samples accordingly;

$$\hat{s}[n] = g[n] + \sum_{k=1}^{P} a_k s[n-k] \tag{2.33}$$

where $\hat{s}[n]$ and $s[n]$ are the estimated and original speech signal, and a_k are the past P linear prediction coefficients. The $g[n]$ is the source signal and can be found by first estimating the prediction coefficients a_k by minimizing the prediction error as shown in Figure 2.12.

$$g[n] = \hat{s}[n] - \sum_{k=1}^{P} a_k s[n-k] \tag{2.34}$$

The prediction coefficients can be estimated by one of above mentioned methods such as linear prediction coding [14], weighted linear prediction [17], its variants extended weighted linear prediction [18], stabilized weighted linear prediction [30], discrete all-pole [19], etc.

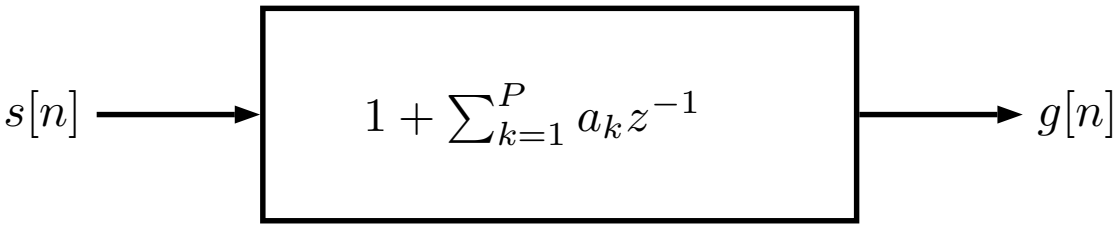

FIGURE 2.12: Linear prediction analysis based source signal

2.2.2.2 Minimum/Maximum phase speech decomposition

The speech is considered as the mixed phase signal, where glottal source signal corresponds to the maximum-phase of signal, and vocal tract is assumed minimum-phase of the speech signal [31–36]. Therefore separating minimum and maximum phase parts of speech give the source and vocal tract filter.

2.2.2.2.1 Zeros of the Z-transform

One of the method to decompose the windowed speech into its maximum and minimum phase is Zeros of the Z-transforms (ZZT) [37]. The z-transform of the windowed speech $s[n]$ is given as [37];

$$S(z) = s[0]z^{-N+1} \prod_{k=1}^{M_i} (z - Z_{C,k}) \prod_{k=1}^{M_0} (z - Z_{AC,k}) \tag{2.35}$$

where Z_C and Z_{AC} are the zeros outside and inside the unit circle respectively. The M_i and M_o are number of inside and outside the unit circle zeros. The method graphically is shown in the Figure 2.13. The roots of the $S(z)$ are classified based on the modulus of the roots. The roots with modulus > 1 are classified as maximum-phase components of speech, and the roots with modulus < 1 are minimum-phase components. Based on the roots spectrum is calculated and then IFFT is taken for the desired component of the speech signal [37].

2.2.2.2.2 Complex cepstrum decomposition

Another methods for maximum-minimum phase decomposition is based on the complex cepstrum of windowed speech $s[n]$ [37], and the Figure 2.14 shows the steps. The complex

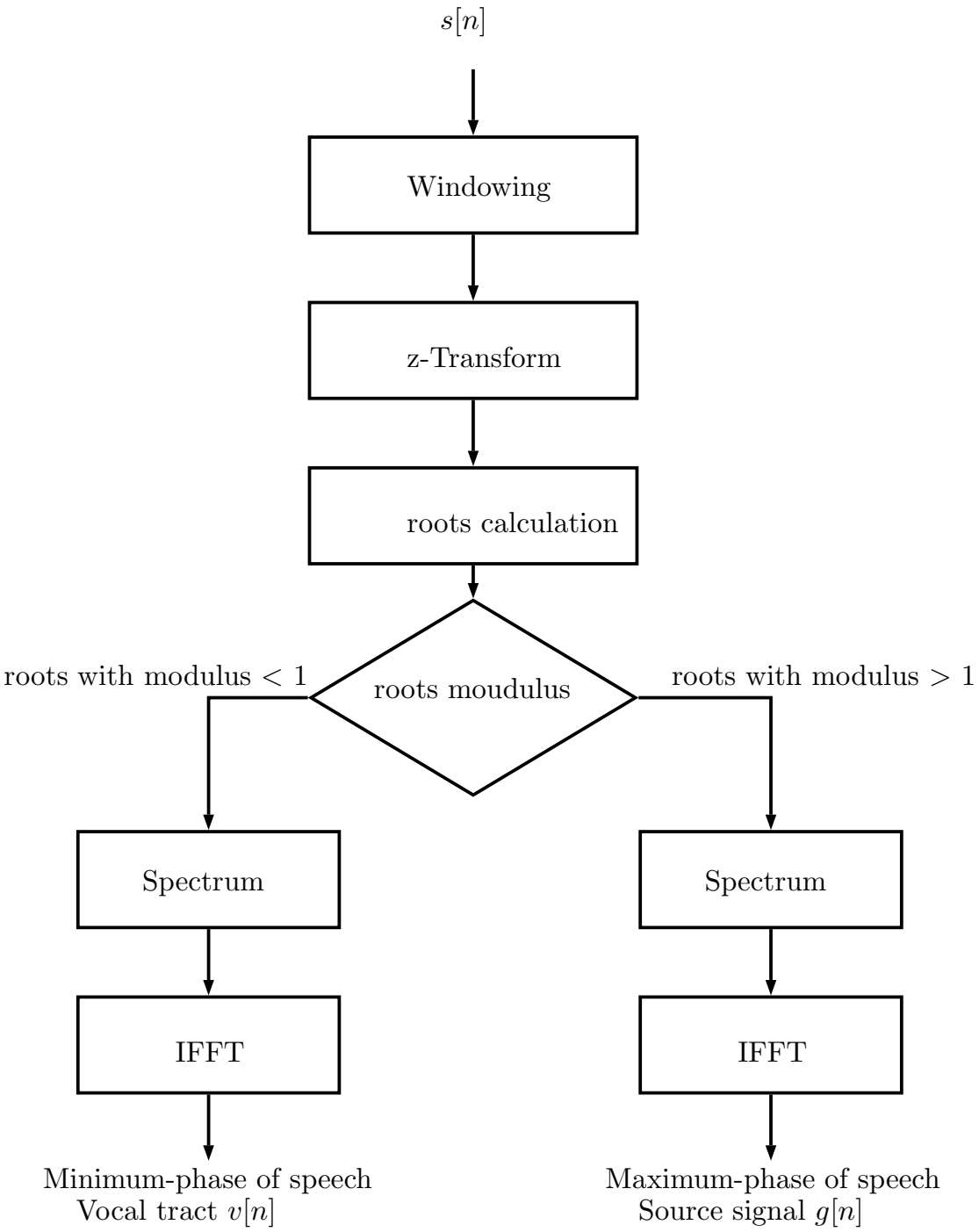

FIGURE 2.13: Zeros of z-transform (ZZT)-based decomposition (adapted from [37])

cepstrum of $s[n]$ is given as [38];

$$S(\omega) = \sum_{n=-\infty}^{\infty} s[n]e^{-j\omega n} \tag{2.36}$$

$$log[S(\omega)] = log(|S(w)|) + j\angle S(\omega) \tag{2.37}$$

$$\hat{s}[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} log[S(\omega)]e^{j\omega n}d\omega \tag{2.38}$$

The flow if the algorithm is as follow, the complex cepstrum can be divided into following components based on the positive and negative index of cepstrum [37];

$$\hat{s}[n] = \begin{cases} |s[0]| & \text{for } n = 0 \\ \sum_{k=1}^{M_0} \frac{Z_{AC,k^n}}{n} & \text{for } n < 0 \\ \sum_{k=1}^{M_i} \frac{Z_{C,k^n}}{n} & \text{for } n > 0 \end{cases} \tag{2.39}$$

The cepstrum correspond to maximum-phase component of signal for negative index of cepsturm, while for positive it corresponds to minimum-phase components of the windowed speech $s[n]$ [37].

2.2.2.3 Iterative Adaptive Inverse Filtering (IAIF)

The IAIF [1] is a highly accurate, and simple Glottal Inverse Filtering (GIF) methods used to decompose the input speech into its source and vocal tract components [4, 39]. The IAIF estimates the vocal tract filter and lip radiation, and then canceled these components from the input for the highly accurate glottal source signal. According to the [39], the input speech signal $s_0[n]$ first highpass filtered with a cutoff frequency less than the fundamental frequency F_0 (i.e. 50 Hz for ES):

$$s[n] = s_0[n] * h_{hp}[n] \tag{2.40}$$

where $h_{hp}[n]$ and $*$ are highpass filter and convolution operator, respectively. Then highpassed filter signal $s[n]$ is inverse filtered with first order all-pole model H_{g_1} for canceling lip radiation and glottal flow from the speech signal:

$$S_{g_1}(z) = \frac{S(z)}{H_{g_1}(z)} \tag{2.41}$$

where $S_{g_1}(z)$, and $S(z)$ are z-transform of $s_{g_1}[n]$ and $s[n]$, respectively. The $s_{g_1}[n]$ further used the all-pole model of order p $H_{vt_1}(z)$ for vocal tract estimation. The $s[n]$ is then

[1] This section mostly taken from [4, 39]

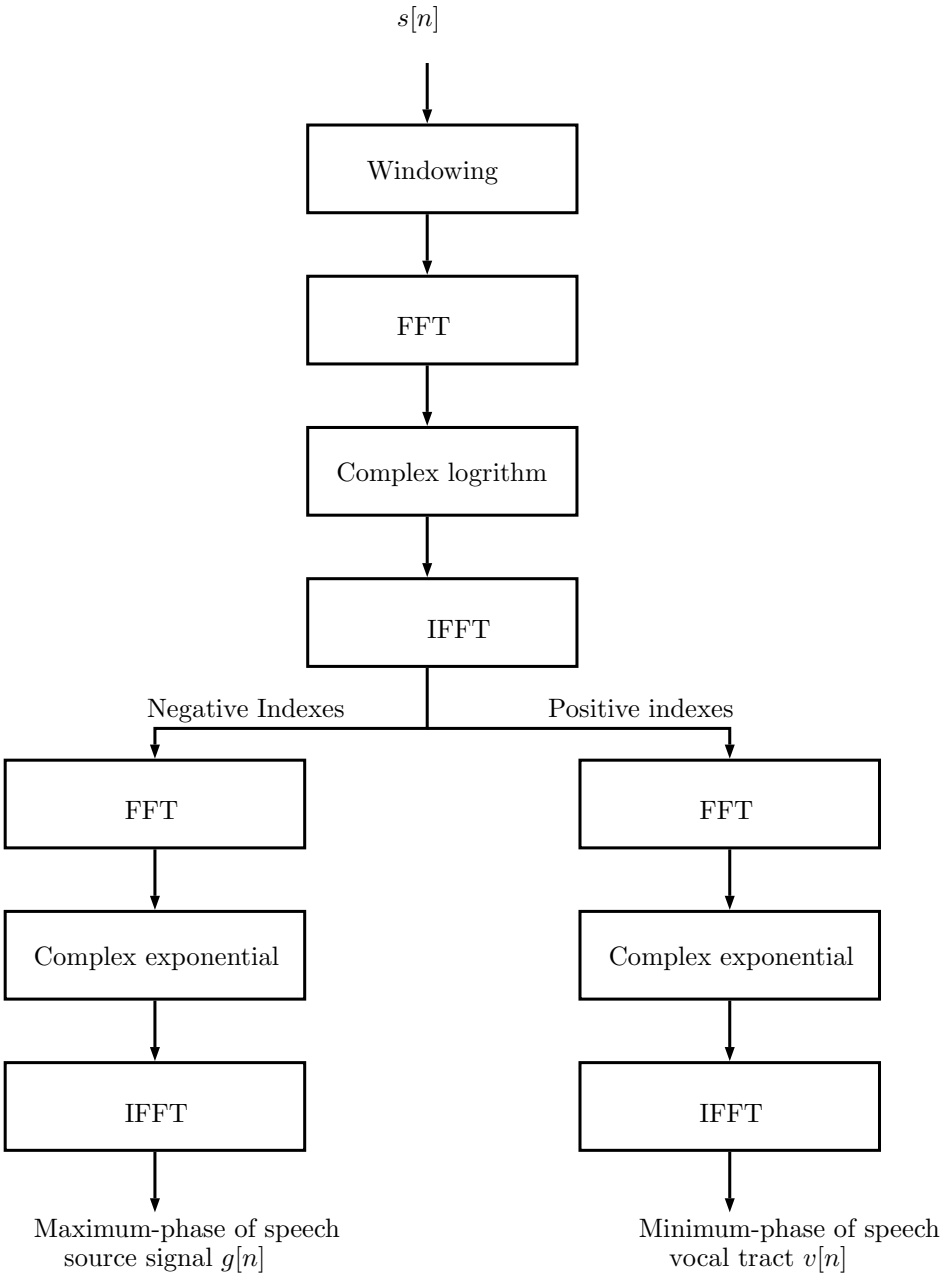

FIGURE 2.14: Complex Cepstrum (CC)-based decomposition (adapted from [37])

inverse filtered with $H_{vt_1}(z)$ for the initial estimate of glottal signal $\hat{g}_1[n]$:

$$\hat{G}_1(z) = \frac{S(z)}{H_{vt_1}} \tag{2.42}$$

where $\hat{G}_1(z)$ is the z-transform of $\hat{g}_1[n]$. The first phase estimation of glottal signal is done by integrating $\hat{g}_1[n]$:

$$g_1[n] = \hat{g}_1[n] + g_1[n-1] \tag{2.43}$$

where $g_1[n]$ is the first estimate of glottal source signal. The 2^{nd} estimate of glottal source spectrum is estimated by all-pole model of order g (typically 2 to 4) $H_{g_2}(z)$. The

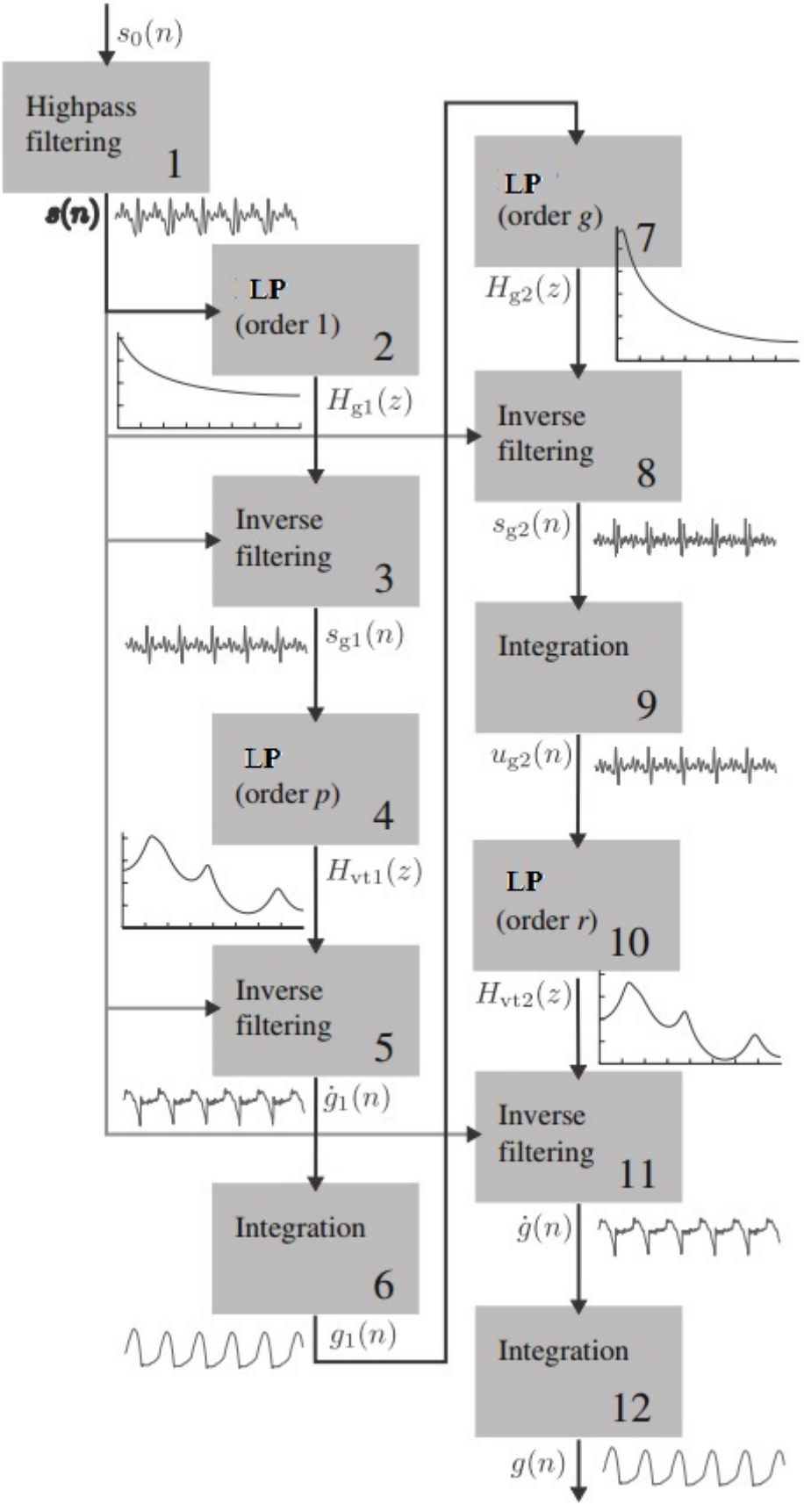

FIGURE 2.15: Iterative Adaptive Inverse Filtering (IAIF) (adapted from [39])

$s[n]$ is then inverse filter with this new all-pole model for glottal flow free signal $s_{g_2}[n]$:

$$S_{g_2}(z) = \frac{S(z)}{H_{g_2}(z)} \tag{2.44}$$

where $S_{g_2}(z)$ is the z-transform of $s_{g_2}[n]$. The lip radiation is then canceled by integrating $s_{g_2}[n]$:

$$u_{g_2}[n] = s_{g_2}[n] + u_{g_2}[n-1] \tag{2.45}$$

where $u_{g_2}[n]$ is the lip radiation and glottal source free signal. The all-pole model (of order p) of $u_{g_2}[n]$ gives the glottal flow derivative signal $\hat{g}[n]$ by inverse filtering the $s[n]$ with $H_{vt_2}(z)$:

$$\hat{G}(z) = \frac{S(z)}{H_{vt_2}(z)} \tag{2.46}$$

where $\hat{G}(z)$ is the z-transform of $\hat{g}[n]$. Finally, the glottal flow $g[n]$ is obtained by integrating $\hat{g}[n]$:

$$g[n] = \hat{g}[n] + g[n-1] \tag{2.47}$$

The $g[n]$ and $H_{vt_2}(z)$ are the glottal source and vocal tract filter of the input speech signal $s[n]$.

2.2.2.4 Closed phase inverse filtering

The Closed Phase Inverse Filtering (CPIF) is another method for decomposing the speech signal into its source and filter components. The idea behind CPIF is to detect the glottal closure and glottal opening instant [40], and then between these instants, there is no interaction between source and filter components, i.e. source signal is absent in this period. The signal in this period only consists of filter components, so the speech signal modeled using Discrete All Pole (DAP) [19, 41], and then inverse filtering the speech signal with DAP gives the desire glottal source, the DAP filter response the desire vocal tract filter [11].

2.3 Acoustic Parameterization

2.3.1 Vocal tract parameterization

The vocal tract $V(z)$, an all-pole model, simply given as;

$$V(z) = \frac{1}{A(z)} = \frac{1}{1 + \sum_{k=1}^{P} a_k z^{-k}} \tag{2.48}$$

where a_k are the linear prediction coefficients of order P, and can be estimated by different prediction coefficients methods mentioned above such LP [14], WLP [17], XLP [18], SWLP [15], and DAP [19] etc. The $A(z)$ is the inverse filter of $V(z)$. The prediction coefficients can be used for further analysis by finding the roots of inverse filter $A(z)$, which normally are the peaks in the prediction spectrum and it represents the poles of vocal tract, and normally can be define mathematically using this simple equation;

$$z_i = r_i e^{j\theta_i} \tag{2.49}$$

where z_i is an ith pole with r_i magnitude and angle θ_i. The poles can be used to find the formants frequencies and formants bandwidth.

2.3.1.1 Formant frequencies

The formant frequencies of the vocal tract calculation require angle of the poles and given as [42, 43]:

$$F_i = \frac{f_s}{2\pi}\theta_i \tag{2.50}$$

where F_i is the formant frequency for the i pole with phase θ_i, and f_s is the sampling frequency. In order to select the accurate estimation, roots solving is used with peak picking from the spectrum of the vocal tract filter, as can be seen in Figure 2.16. Some of the roots does not correspond to the formants, i.e. formants has the radius less than the 0.90 on the unit circle, as can be seen from Figure 2.16. The formant at 217 Hz does not correspond to the formant frequency as its radius on unit cirle is less than 0.90, so it is discarded and shown by cross on it.

2.3.1.2 Formant Bandwidths

The bandwidths of the formant frequencies is calculate from the poles according to following mathematical equation [42, 43]:

$$B_i = -\frac{f_s}{\pi} ln(r_i) \tag{2.51}$$

where r_i is the magnitude of pole i, f_s is the sampling frequency of the signal, and B_i is the 3-dB bandwidth of formant i. The Figure 2.16 shows the bandwidths of the formants.

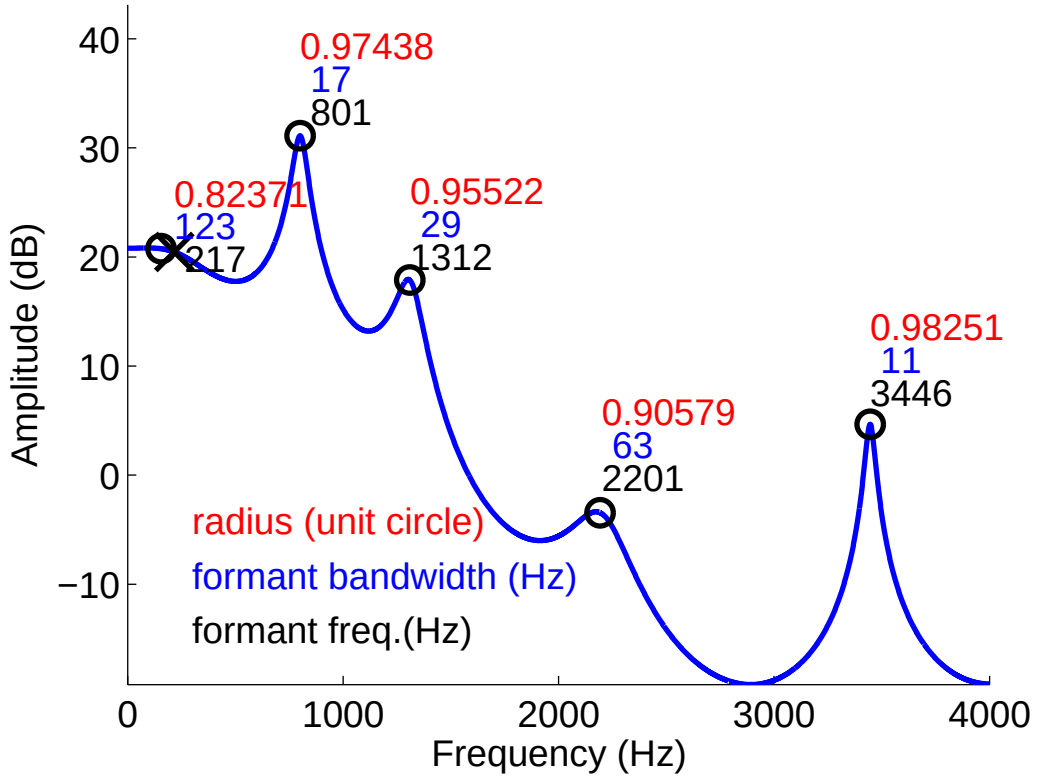

FIGURE 2.16: Formant frequencies and corresponding bandwidths for Spanish vowel /a/

2.3.1.3 Vocal tract spectrum

The spectra of vocal tract filter coefficients is computed using:

$$H(e^{j\omega}) = 20log|\frac{1}{A(e^{j\omega})}| \tag{2.52}$$

where $A(e^{j\omega})$ is given in Equation 2.48, and spectra is shown in the Figure 2.16.

2.3.2 Source signal parameterization

The source signal an important component of speech production can be further decomposed or parameterized into different components, which convey different information regarding the source signal.

2.3.2.1 Time domain parameters

To exploit the pseudo-periodicity of the voiced speech for pitch synchronous processing of the speech has advantages in many fields of speech such as speech enhancement, speech

recognition, speech coding etc. Figure 2.17 shows the instants of glottal closure. The difference between subsequent GCIs gives pitch period of the signal. The Glottal Closure

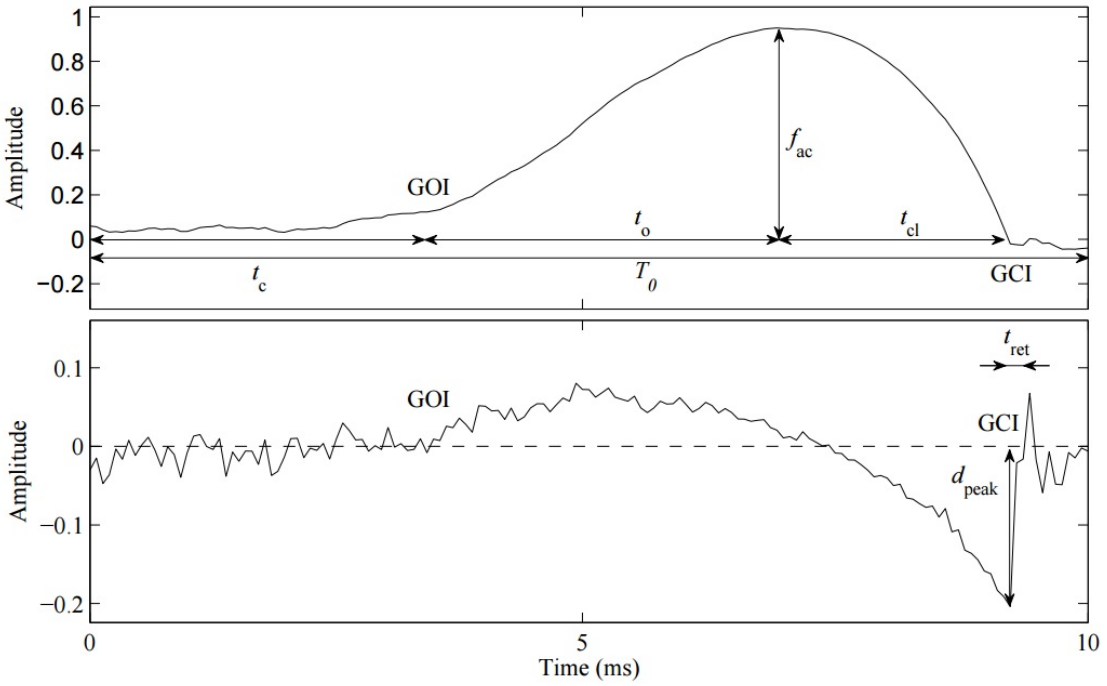

FIGURE 2.17: Glottal Closure Instant from the signal cycle of source signal with its derivative (adapted from [44])

Instants (GCIs) are used to find the pitch contours of speech, as well the boundaries of individual source cycle for much and better analysis and synthesis, such as source signal estimation [45], prosodic modification of speech [46], speech dereverberation [47], and speech synthesis [48] etc. There are some methods available in literature for GCI, such as Hilbert envelope of source signal based [49], the Dynamic Programming Phase Slope Algorithm (DYSPA) [50], group delay based estimation [51], Zero Frequency Resonator (ZFR) [52], Yet Another GCI Algorithm (YAGA) [53], and the most recent one is Speech Event Detection using the Residual Excitation and a Mean-based Signal (SEDREAMS) [54].

The other time domain parameters which are estimated by timing instants of singla glottal pulse as shown in Figure 2.17 are Open Quotient (OQ) [55], Speech Quotient (SQ) [55], Closing Quotient (CIQ) [56], and amplitude based is Normalized Amplitude Quotient (NAQ) [57].

$$OQ = \frac{(t_0 + t_{cl})}{T_0} \tag{2.53}$$

$$SQ = \frac{t_0}{t_{cl}} \tag{2.54}$$

$$CIQ = \frac{t_{cl}}{T_0} \tag{2.55}$$

$$NAQ = \frac{f_{ac}}{d_{peak} T_0} \tag{2.56}$$

where t_0, t_{cl}, T_0 are opening timing, closing time, and fundamental period respectively. The maximum value of source flow signal is f_{ac} and minimum of source flow derivative is d_{peak}. These time domain parameters provide different aspects of source signal ,and used for different purposes in the speech signal processing algorithms.

2.3.2.2 Frequency domain parameters

The frequency domain parameters of source signal also provide information about the type of phonation, about its harmonicity, and its periodicity. Based on the spectrum of source signal shown in Figure 2.18, H1-H2 difference i.e. difference between first and second harmonics amplitude is a simple spectral decay measure of speech signal [58]. Another parameter for glottal source is Harmonic Richness Factor (HRF) used to quantify the spectral decay of the speech signal [59], and defined as;

$$HRF = \frac{\sum_{i=2}^{N_{hr}} H_i}{H_1} \tag{2.57}$$

where H_1 is the fundamental harmonic amplitude and H_i are next harmonics amplitudes after fundamental harmonic, and i is the harmonic index. To model the spectral slope, parabolic spectral parameter (PSP) [60], and linear regression have been proposed [61, 62]. The modeling of source spectrum with all-pole model for spectral decay is another simple measure [63]. Besides the spectral delay of the source signal, the source

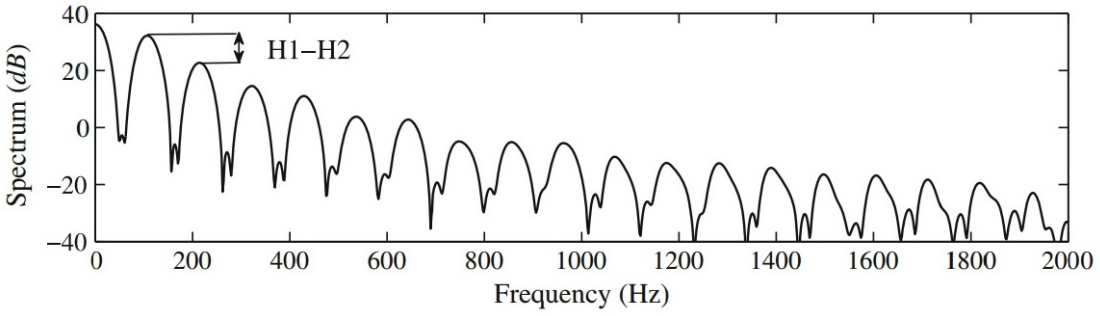

FIGURE 2.18: Frequency spectra of source signal (adapted from [57])

signal has periodic and aperiodic components, and useful for speech modification, and synthesis. The simplest measure for source signal quantization into periodic and aperiodic components is the Harmonic to Noise Ratio (HNR). The HNR is defined as the ratio of periodic and aperiodic components of source signal. The periodic and aperiodic components can be estimated by different methods, based on the each spectral bands of the source spectrum such as the author in [64], calculated correlation coefficients for

several spectral bands, which defined the aperiodic components in each spectral band. Another method for HNR estimation is based on upper and lower smoothed spectral envelopes [65]. The spectral envelops are defined as the amplitude of harmonics peaks and inter-harmonics valleys in the source spectrum. The ratios are then averaged across spectral bands with rectangular bandwidth (ERB) scale [66]. The HNR is normally used to measures the degree of voicing in the speech, and as well to controlled the voiced quality by adjusting HNR in the speech signals [65].

In the previous studies of ES enhancement, linear prediction analysis-synthesis is mainly used. The linear prediction is an efficient and mostly used method in speech signal processing, which parameterize the speech signal into its source and filter components [14]. Linear prediction is based on the concept of speech as an output of Linear Time Invariant (LTI) system, whose input is the periodic or noisy source based on type of phonation [3]. Previous studies, also considered this fact, and assume that ES can be faithfully modeled using its source and filter components, and it has been used extensively. In comparison to normal speech, the pharyngo-esophagus segment vibration provides the ES source, and vocal tract represent ES filter part.

Initially, the analysis of ES studies have provided, the essential characteristics of corresponding source and filter components. Some of these studies provided different aspects of ES, such as [67] provided the comparison of source signal of ES with normal speech based on averaging, and concluded that ES has low irregular fundamental frequency or no periodicity. The source correspond whispered speech source signal. The analysis of vocal tract filter of the sustained vowels have revealed that the spectral peaks (formants) of ES are shifted upward in the frequency [68]. The quality rating analysis of ES in comparison to normal and artificial larynx speech has been the subject of the [69]. ES source signal characteristics revealed high jitter and roughness in the study [70]. The authors in [71] studied the long time spectral and intensity characteristics, and compare it with normal speech for the differences. An extension to the previous study of ES [71], authors further characterized the acoustic and temporal of the ES [72] and also provided differences between TES, ES and normal speech [73, 74]. During these years one the of the analysis study about spectral and temporal characteristics provided the baseline for future studies [75, 76]. Some of the other analysis studies, commonly used in ES research community have provided the detail analysis along with synthesis of the ES for better and intelligible ES are presented in [77–79].

Using the source-filter decomposition by linear prediction analysis-synthesis, Qi in [77] used the vowels /i/, /a/, /e/, /u/ and one diphthong /ou/. The vocal tract of the vowels is modified by removing real poles, and then synthesis with original source signal and modified vocal tract for enhanced version of ES. The modification to [77], is done

by replacing the source signal with synthetic source signal. The synthetic source signal was obtained using LF source model. The results thus obtained with this system, has good intelligible speech, but it sounds robotics. The source signal has been modified by first estimating the fundamental frequency using the autocorrelation function, and then modifying fundamental frequency by smoothing it, and using it with LF source model, with the original vocal tract for synthesis [80]. The results were promising in comparison to previous methods, but still it sounds more robotic due to synthetic source model. The fundamental frequency smoothed curve based source signal was synthesized with original vocal tract for more intelligible ES in [81]. The vocal tract spectral peaks (formants) of ES vowels has been smoothed and then synthesized with original source signal, the results showed significant perceptual enhancement [82]. The bandwidth of the spectral peaks (formants) has been increased along with LF source model based source signal for synthesis for better and intelligible ES [83]. Perceptual enhancement was obtained by introducing radiated pulses in the source signal in frequency domain [84]. Another method has much better intelligible ES by applying comb filtering to source signal by introducing excitation instants to the source signal [81]. The statistical transformation also has been used in literature, by transforming ES into normal speech, but it needs lot of speech samples [85, 86]. The modification to statistical transformation of ES to normal speech using the eigenvoices, has suboptimally enhanced the intelligibility of ES [87]. The use of Kalman filtering [88–97] for enhancing the vocal tract coefficients has significant improvement. The Kalman filter, and the poles stabilization method [98–101] has provided much better intelligible ES. The use of Kalman filtering in modulation domain [102–109] also has shown promising results [110, 111] in a sense of improved HNR [112, 113]. The recognition based system which replaced the recognized vowels with normal speech vowels for better and perceptual ES [114]. A software called ESOIMPROVE has been developed by Deusto university which provided significant freedom to analysis and synthesis the ES, and observe the results in a sense of Harmonic to Noise Ratio (HNR), Jitter, Shimmer [115]. A system is developed using the Kalman filtering applied to wavelet transformed based signal, and using the pole stabilization and the results obtained using Multi-Dimensional Voice Program (MDVP) has significant enhancement in HNR, and perception [100]. The estimation of periodicity of the source cycle for accurate pitch, shimmer and jitter for better and enhanced version of ES has provided perceptually better ES [116]. A vector quantization based speech conversion used to modify the vocal tract of ES for better sounding ES[117]. Some of the other, enhancement method and as well the analysis studies are [118–165]. The complete list of studies about ES is shown in table 2.1.

Study	Analysis		Synthesis		Method	Year
-	Source	Filter	Source	Filter	-	-
Weinberg and Bennet [67]	√	x	x	x	LPC	1972
Sisty and Weinberg [68]	x	√	x	x	LPC	1972
Bennet and B. [69]	√	√	x	x	LPC	1973
Smith and Horri [70]	√	x	x	x	LPC	1978
Weinberg and Smith [71]	√	√	x	x	LPC	1980
Weinberg [72]	√	√	x	x	LPC	1982
Robbins and Singer [74]	√	√	x	x	LPC	1984
Weinberg [73]	√	√	x	x	LPC	1986
Nord and Hammarberg [75]	√	√	x	x	LPC	1989
Trudea and Qi [76]	√	√	x	x	LPC	1990
Yingyong et al. [77]	√	√	√	x	LPC	1995
Tull and Rutledge [81]	√	x	x	x	LPC	1993
Qi [78]	√	√	√	x	LPC	1995
Qi and B. [79]	√	x	x	x	LPC	1995
Kenji and Noriyo [82]	x	√	x	√	LPC	1999
Cervera et al. [166]	x	√	x	x	LPC	2001
Prosek and Vreeland [87]	x	x	x	√	LPC	2001
Kenji et al. [125]	x	√	x	√	LPC	2002
Garcia et al. [115]	√	x	x	x	LPC	2005
Loscos and Bonada [84]	x	√	x	√	LPC	2006
Ali and Jebara [83]	√	√	√	√	LPC	2006
Alfredo et al. [114]	x	√	x	√	LPC/Neural network	2006
Sirichokswad et al. [80]	x	√	√	√	LPC	2006
Garcia et al. [98]	x	√	x	x	Wavelet	2006
Garcia and Mendez [99]	x	√	x	√	LPC/Kalman	2008
Garcia et al. [116]	√	x	x	x	LPC	2009
Sabayjai et al. [86]	x	√	x	√	LPC	2009
Doi et al. [85]	√	√	√	√	GMM	2010
Sharifzadeh et al. [167]	√	√	√	√	CELP	2010
Ibon et al. [100]	x	√	x	√	LPC/Kalman	2010
Isasi et al. [101]	x	x	√	x	LPC	2011
Ferrat and Guerti [168]	√	√	x	x	LPC	2012
Ishaq and Zapirain [162]	√	x	√	√	LPC	2012
Ishaq et al. [110]	√	√	√	√	LPC/Kalman	2013
Ishaq and Zapirain [111]	√	√	√	√	LPC/Kalman	2014
Mcloughlin et al. [169]	√	√	√	√	CELP	2015

TABLE 2.1: Studies on Esophageal Speech (ES)

2.4 Chapter Summary

This chapter of the thesis has provided a review for the normal, and pathology speech production methods. The source filter model of the human speech for linear time-invariant system described, with reference to detailed components of source and vocal tract components. The source and vocal tract filter modeling and parameterization has given in detail. The source-filter decomposition methods has provided. At the end the previous work related to ES enhancement has elaborated. The techniques presented in this chapter are the bases for the rest of thesis for developing the system for ES enhancement.

Chapter 3

System Design

This chapter of thesis provides the proposed system in detail in terms of following essential components, i) Analysis, ii) Transformation, and iii) Synthesis, shown in Figure 3.1. The purpose of the analysis is to divide the input speech $s[n]$ into frames, and subsequently classify each frame as voiced and unvoiced. Each voiced frame is then decomposed into source and filter components. For decomposition, the inverse filtering method is used. An automatic inverse filtering method Iterative Adaptive Inverse Filtering (IAIF) is used for this purpose. The IAIF first estimate the vocal tract and lip radiation, and then iteratively cancel from the speech signal for source signal [4]. After decomposition, the source and vocal tract components are transformed into normal speech source and filter components using the natural glottal pulse for the source signal, and for the vocal tract using the second order Frequency Warping Function (FWF). The second order FWF moves the formants to lower frequency as they are moved to higher frequency in Esophageal Speech (ES). The source signal is parameterized into Harmonic to Noise Ratio (HNR), fundamental frequency F_0 and the source signal spectrum $G(z)$. The transformation transforms the source and filter components into normal speech components using natural glottal pulse and second order Frequency Warping Function (FWF). Using these source parameters, the source signal transformation, uses any arbitrary normal speech F_0 curve. The natural glottal pulse is then interpolated using FFT based interpolation method using the F_0^N. The gain, HNR, spectral and lip radiation are then compensated for enhanced source signal. The vocal tract filter is modified using the formant frequencies and bandwidth smoothing, and then shifting the spectral peaks to the lower frequency using FWF. The bandwidth of formants are broaden using widening filter. The modified source and filter components are then synthesized for the enhanced version of ES.

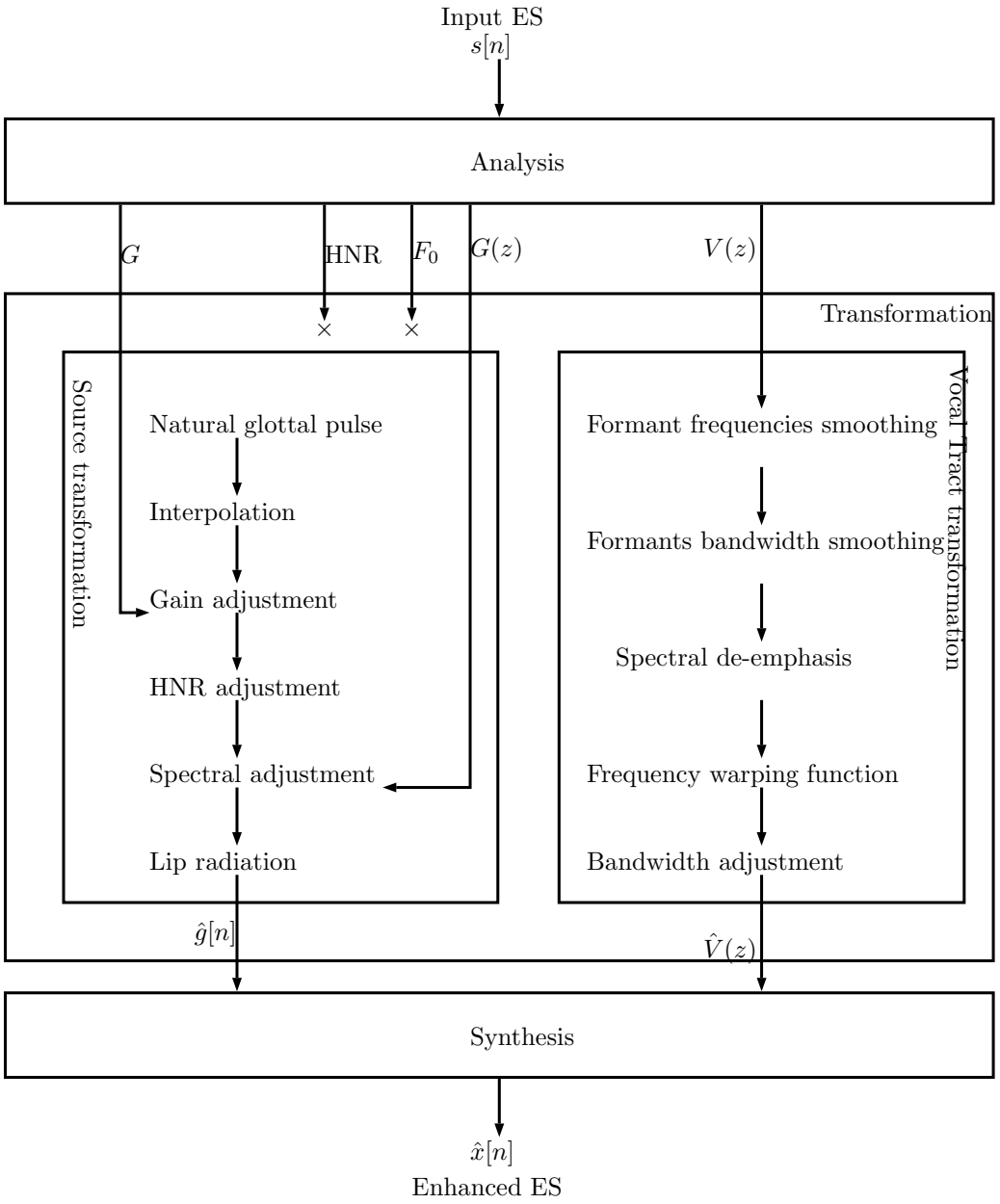

FIGURE 3.1: Proposed enhancement system

3.1 Analysis

The purpose of the analysis is to decomposes the frames into source and filter components, and shown in Figure 3.2. An automatic inverse filtering method Iterative Adaptive Inverse Filtering (IAIF) [4] is used for decomposition. After decomposition, the analysis part, parameterize the source and filter into further subcomponents. The source is parameterized into Harmonic to Noise Ratio (HNR), fundamental frequency F_0, and source spectrum $G(z)$. The vocal tract is transformed into its spectrum $V(z)$.

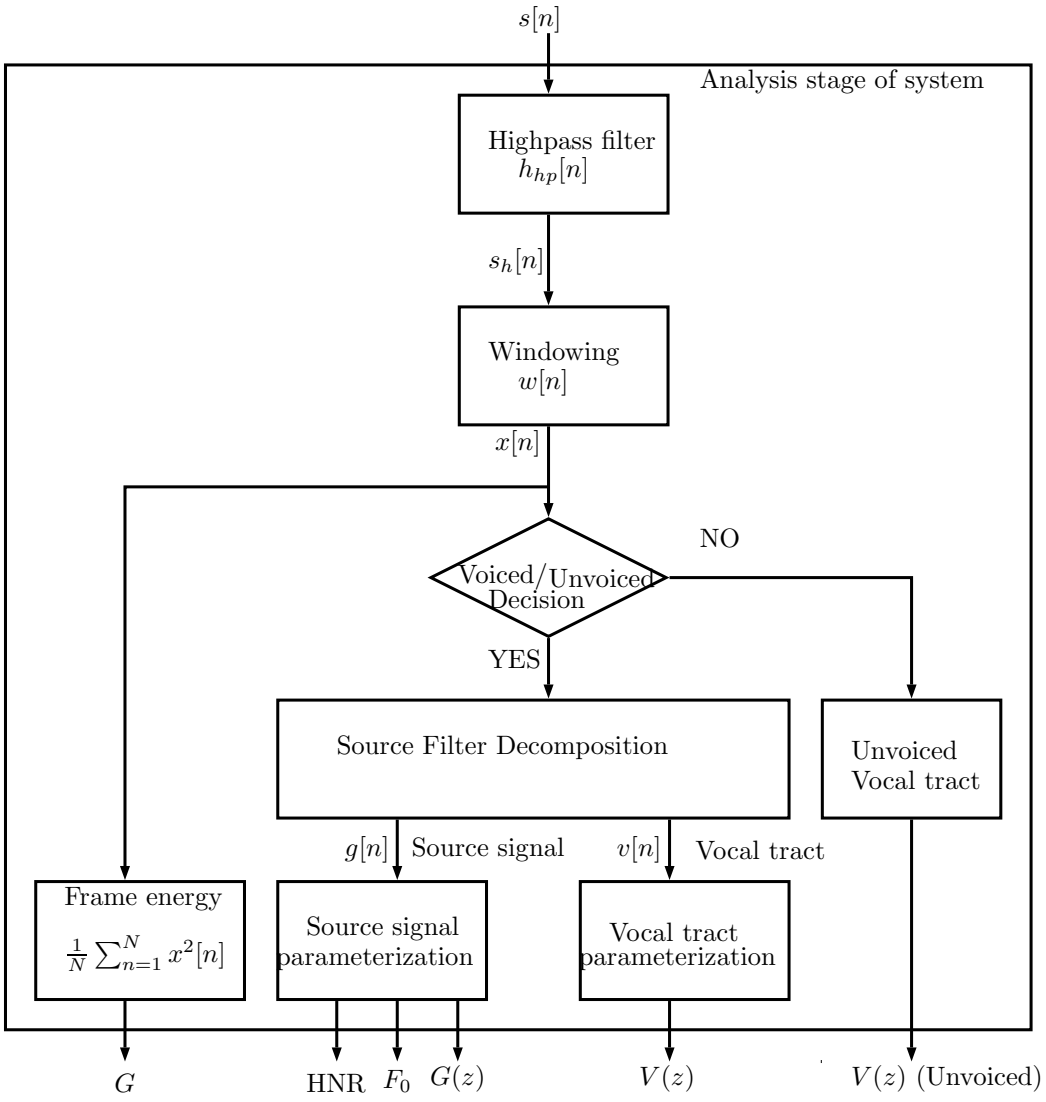

FIGURE 3.2: Analysis part of proposed enhancement system

3.1.1 Highpass filtering

The input ES signal $s[n]$ is first passed through a finite impulse response highpass filter $h_{hp}[n]$ with a cutoff frequency 50 Hz and order of 300 taps, to reduces the low frequency fluctuation;

$$s_h[n] = s[n] * h_{hp}[n] \tag{3.1}$$

where $s_h[n]$ is a highpass filtered signal and $*$ is a convolution operator. The least square linear-phase finite impulse response filter with 300 taps is used in the system. The Figure 3.3 shows the signal before and after the filtering.

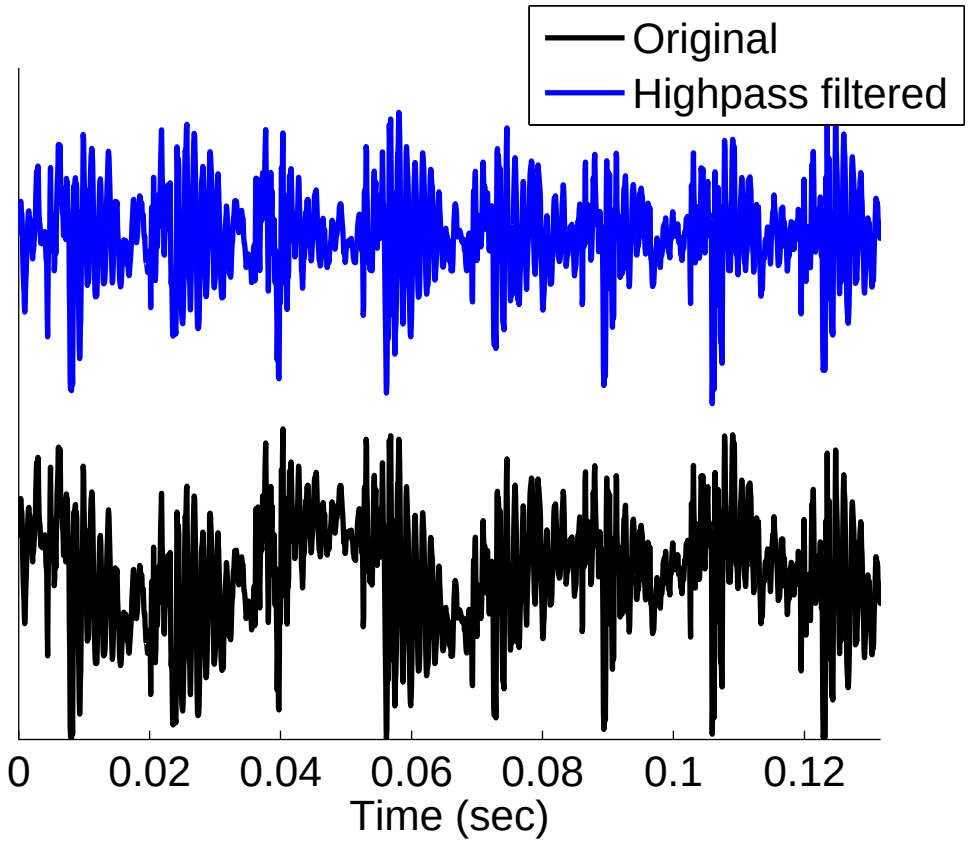

FIGURE 3.3: Highpass filtered signal with original signal

3.1.2 Windowing

The highpass filter signal $s_h[n]$ is then divided into frames using the Hanning window of size 30-ms with 5-ms shift;

$$x[n] = s_h[n].w[n] \tag{3.2}$$

where $x[n]$ is a windowed frame, and $w[n]$ is a Hanning window (shown in Figure 3.4), and given as;

$$w[n] = \begin{cases} 0.5(1 - cos[2\pi \frac{n}{N}]) & 0 \leq n \leq N \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

where N is number of samples in a window. Other windows such as rectangular, hamming, and kaiser can be used, but it is recommend from speech processing research that Hanning window is the most optimal option.

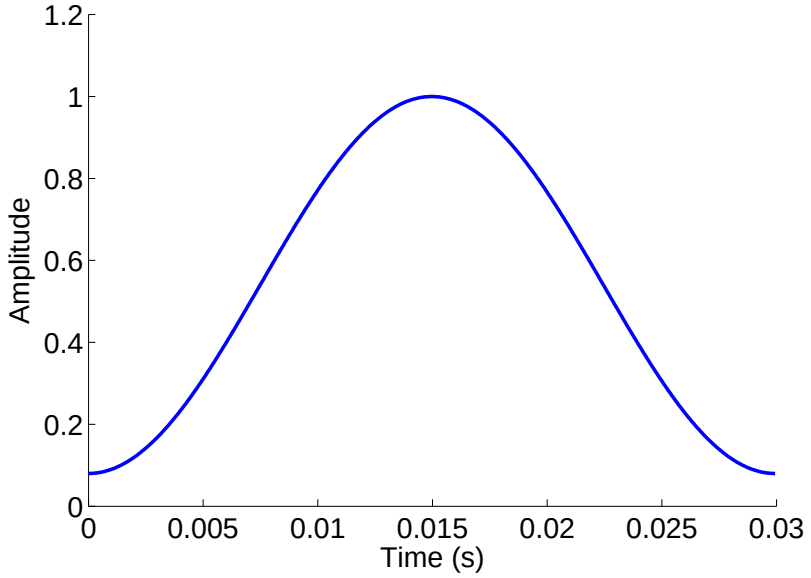

FIGURE 3.4: Hanning window of length 30-ms

3.1.3 Frame energy

After windowing the speech signal, energy of each frame is calculated;

$$G = \frac{1}{N} \sum_{n=1}^{N} x^2[n] \tag{3.4}$$

where G is the energy of the frame $x[n]$. The Figure 3.5 shows the energy curve for the speech signal with the frame length of 30-ms.

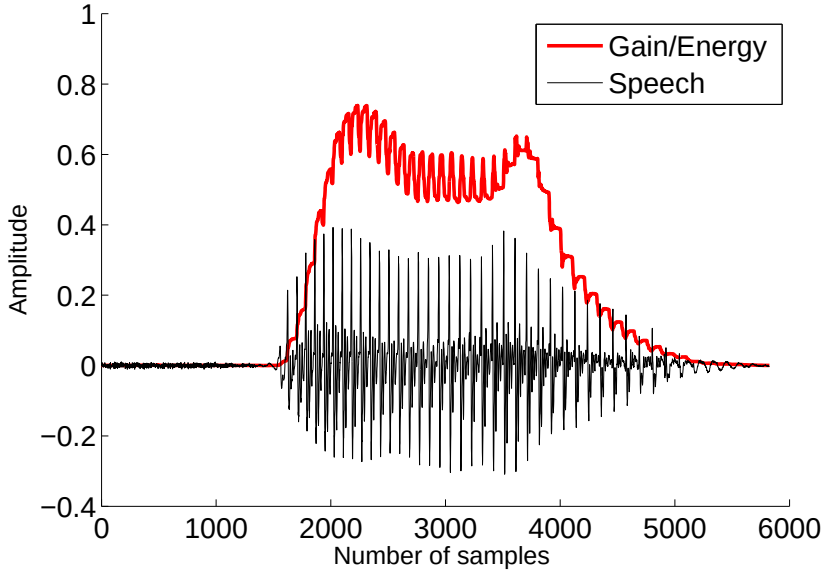

FIGURE 3.5: Energy gain for the speech signal with frame size of 30-ms.

3.1.4 Voiced/Unvoiced decision

The frames are classified into voiced and unvoiced frames. The voiced/unvoiced decision is based on, following measurements, i) energy of frame G, and ii) number of zero-crossing. The decision is shown in Figure 3.6.

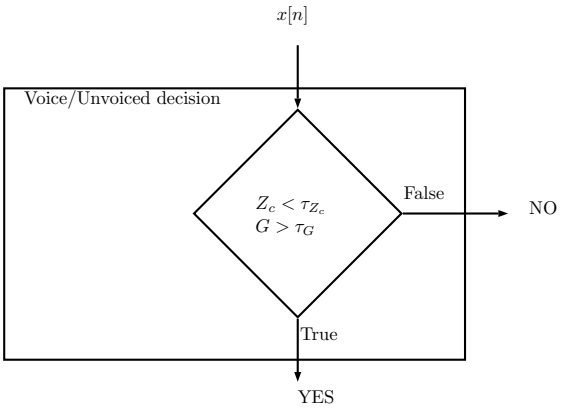

FIGURE 3.6: Voiced/Unvoiced decision

3.1.4.1 Zero crossing

The zero-crossing Z_c of a frame is given as;

$$Z_c = \frac{1}{N}\sum_{n=1}^{N}\frac{|sign\{x[n]\} - sign\{x[n-1]\}|}{2} \tag{3.5}$$

where N is the number of samples in the frame. The zero crossing for voiced frame is low, and for unvoiced its value is high as shown in Figure 3.7. It can be seen from the Figure 3.7, that zero-crossing for voiced speech is low, and for the unvoiced speech, it is high.

Based on these measurements, the voiced/unvoiced decision is made accordingly;

$$VUV = \begin{cases} 1 & \text{if } G > \tau_G \text{ and } Z_c < \tau_{Z_c} \\ 0 & \text{otherwise} \end{cases} \tag{3.6}$$

where τ_G, τ_{Z_c} are the threshold values for energy and zero-crossing for voicing unvoiced decision. Figure 3.8 shows the voiced/unvoiced area of the speech signal.

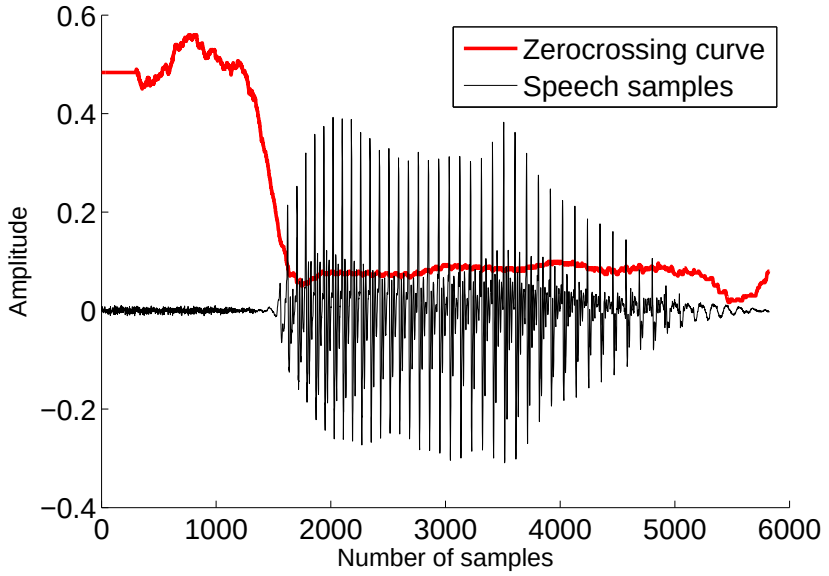

FIGURE 3.7: Zero-crossing for the speech signal with frame size of 30-ms

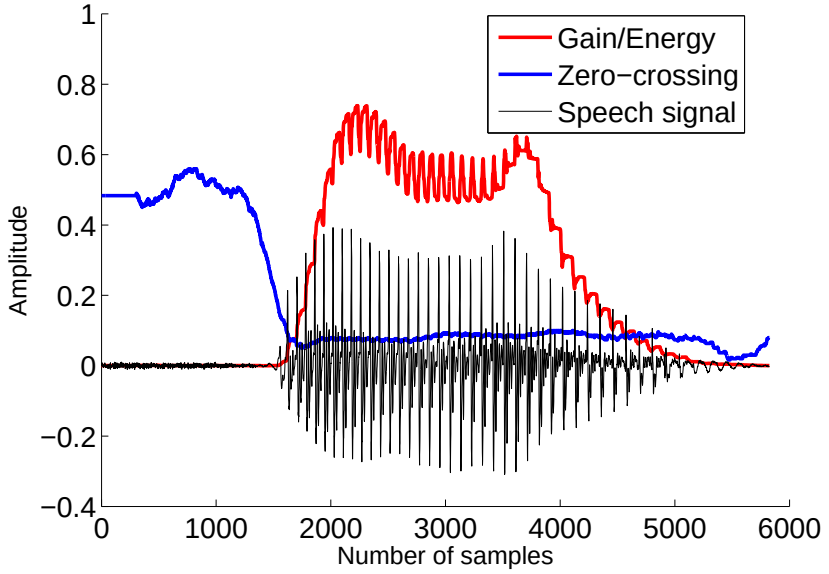

FIGURE 3.8: Zero-crossing and energy/gain for the speech signal with frame size of 30-ms

3.1.5 Source-Filter decomposition

After voiced unvoiced classification, the voiced frame are decomposed into source and vocal tract filter components. The automatic inverse filtering method Iterative Adaptive Inverse Filtering (IAIF) [4] is used for this purpose and shown in Figure 3.9. The IAIF

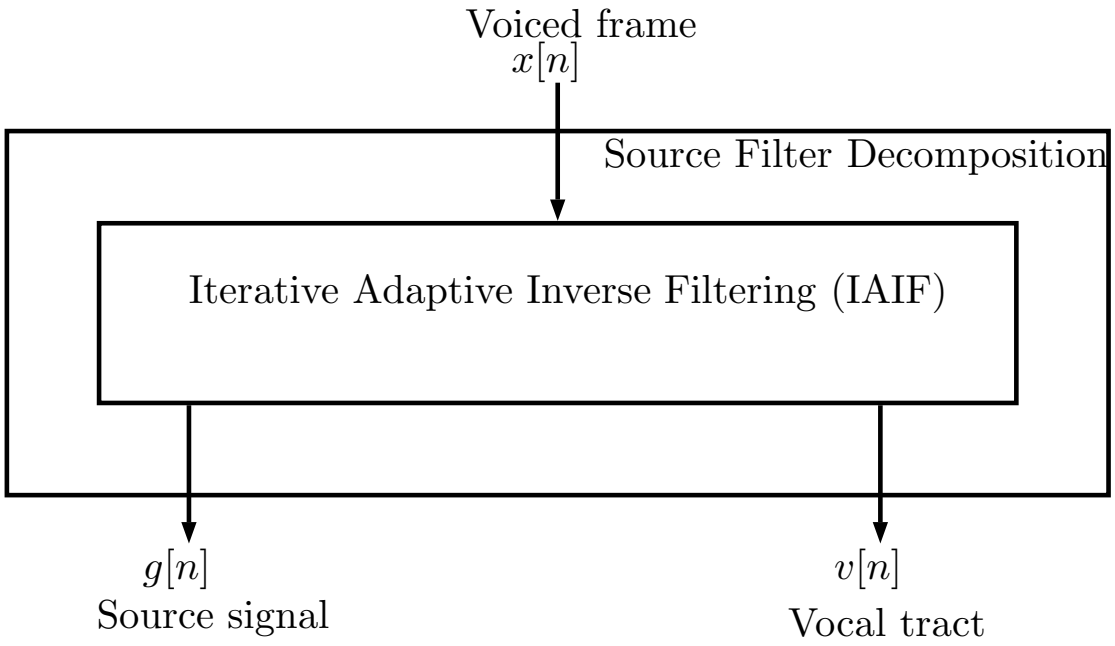

FIGURE 3.9: Source filter decomposition of voiced frame

estimates the lip radiation and vocal tract filter by all-pole model, and then cancel these components from the speech signal iteratively for a source signal [4]. In simplify term;

$$G(z) = \frac{X(z)}{V(z)R(z)} \tag{3.7}$$

where $V(z)$ and $R(z)$ are vocal tract and lip radiation transfer functions, and given as:

$$V(z) = \frac{1}{1 + \sum_{k=1}^{P} a_k z^{-k}} \tag{3.8}$$

where a_k are linear prediction coefficients of order P, and

$$R(z) = 1 - \alpha z^{-1}, \quad 0.96 < \alpha < 1 \tag{3.9}$$

where α is a lip radiation constant. Figure 3.10, and 3.11, shows the source signal and vocal tract filter by IAIF. The vocal tract filter shown in frequency domain is clearly discard the effect of source signal from the spectrum as there is no spectral peaks in lower frequency, which are normally because of source signal effect.

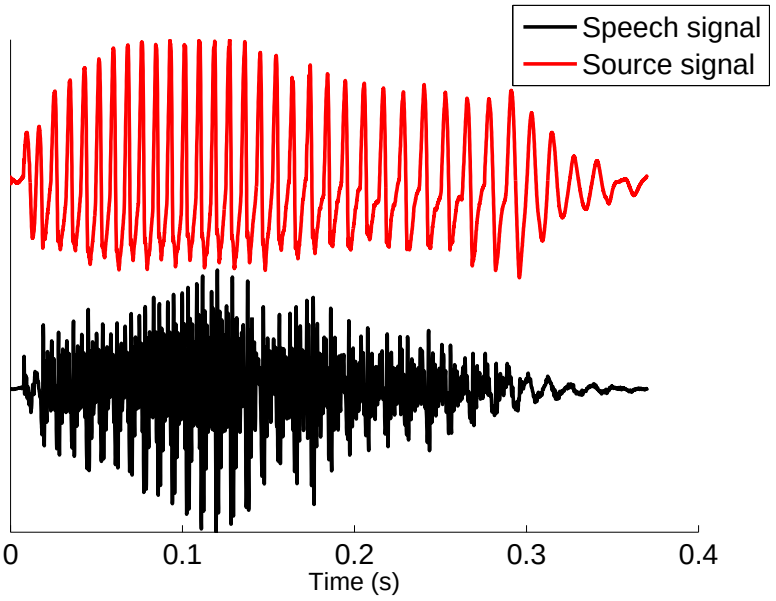

FIGURE 3.10: Source signal obtained by IAIF (vowel /a/)

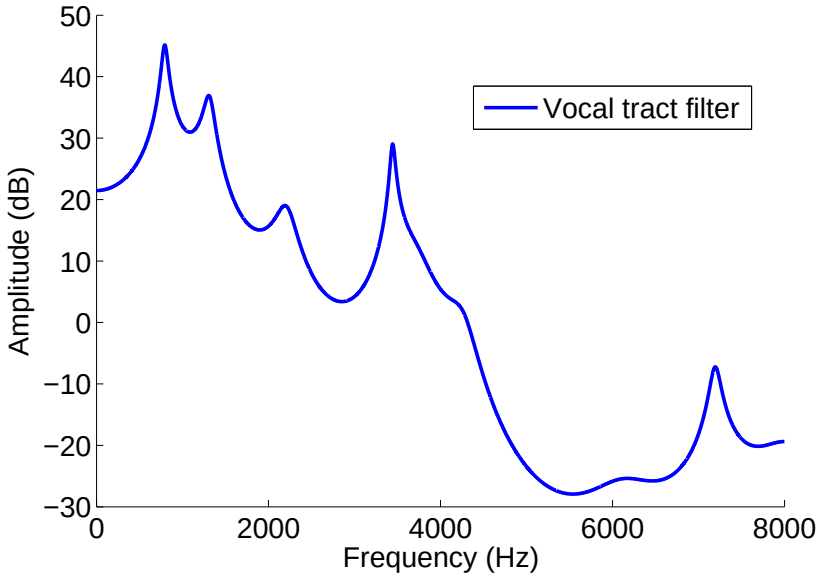

FIGURE 3.11: Vocal tract transfer function obtained by IAIF (vowel /a/)

Once, the input speech signal $x[n]$ is decomposed into source signal $g[n]$ and vocal tract $V(z)$, using IAIF, it can be analyze independent of each other. Both source and vocal tract are then further parameterized into different parameters for analysis and transformation purposes. The source signal is parameterized into fundamental frequency F_0, HNR, and source spectrum. The vocal tract is parameterized into its vocal tract frequency spectrum, showing the spectral peaks as formants, and its width as formants bandwidth.

3.1.6 Source signal parameterization

The source signal $g[n]$ obtained using IAIF is further parameterized into F_0, Harmonic to Noise Ratio (HNR), and its spectrum $G(z)$.

3.1.6.1 Fundamental frequency

The fundamental frequency is estimated by applying the LP analysis with the frame length of 3-ms and overlap of 1 sample. The error variance of LP analysis is calculated for each frame according to:

$$\epsilon(i) = r_{gg}^{i}[n] - \sum_{k=1}^{P} a_k^i r_{gg}^i[k], \quad i = 1, 2, \ldots, N \tag{3.10}$$

where $r_{gg}^{i}[n]$ is the autocorrelation function for frame i, and N is number of frames. The autocorrelation function is given as:

$$r_{gg}[\tau] = \sum_{n=1}^{L} g[n]g[n-\tau] \tag{3.11}$$

where L is number of samples in a frame. The peaks of Error Variance signal (EVS) ϵ corresponds to the excitation instants. The distance between instants gives pitch period of the signal, and inverse of pitch period corresponds the F_0.

3.1.6.2 Harmonic to Noise Ratio (HNR)

The Harmonic to Noise Ratio (HNR) is used to measure the periodicity in the source signal [170]. The evaluation of the HNR follows these simple steps for four frequency bands:

- taking fast Fourier transform (FFT) of windowed speech
- estimating the cepstrum for each frequency band
- for each band HNR is calculated as the ratio of maximum value of cepstral peak to the other quefrencies average value [171]

3.1.6.3 Source spectrum $G(z)$

The spectral tilt of source signal is captured using the LP analysis of order 10 of the source signal.

$$G(z) = \frac{1}{1 + \sum_{k=1}^{P} \alpha_k z^{-k}} \tag{3.12}$$

where α_k is LP coefficients of order $P(10)$ for the source signal $g[n]$.

3.1.7 Vocal tract parameterization

The vocal tract transfer function is then further represented by its spectra.

$$V(e^{j\omega}) = V(z) = 20log|\frac{1}{A(e^{j\omega})}| = 20log|\frac{1}{A(z)}| \tag{3.13}$$

In short the analysis stage of ES provides us the following parameters for each voiced frame:

- frame energy G
- frame vocal tract filter coefficients (order 30) and its spectra $V(z)$
- frame source signal $g[n]$
- frame source spectrum $G(z)$ (order 10)
- frame HNR
- frame F_0

3.2 Transformation

The transformation of the system transform the source signal $g[n]$ of ES into normal speech and vocal tract $v[n]$ to nearly normal speech vocal tract. It is compulsory to analysis source and vocal tract before actual transformation applied, so reason behind this transformation can be justified. Analyzing the source signal obtained from sustained Spanish vowel (/a/, /e/, /i/, /o/, /u/), it can be observed from the HNR, that the ES lacks periodicity in comparison to normal speech, as shown in Figure 3.12. The time domain analysis of source signal reveals that their is no periodic components, and it can be observed clearly in Figure 3.13. The spectral tilt of source signal also does not follow

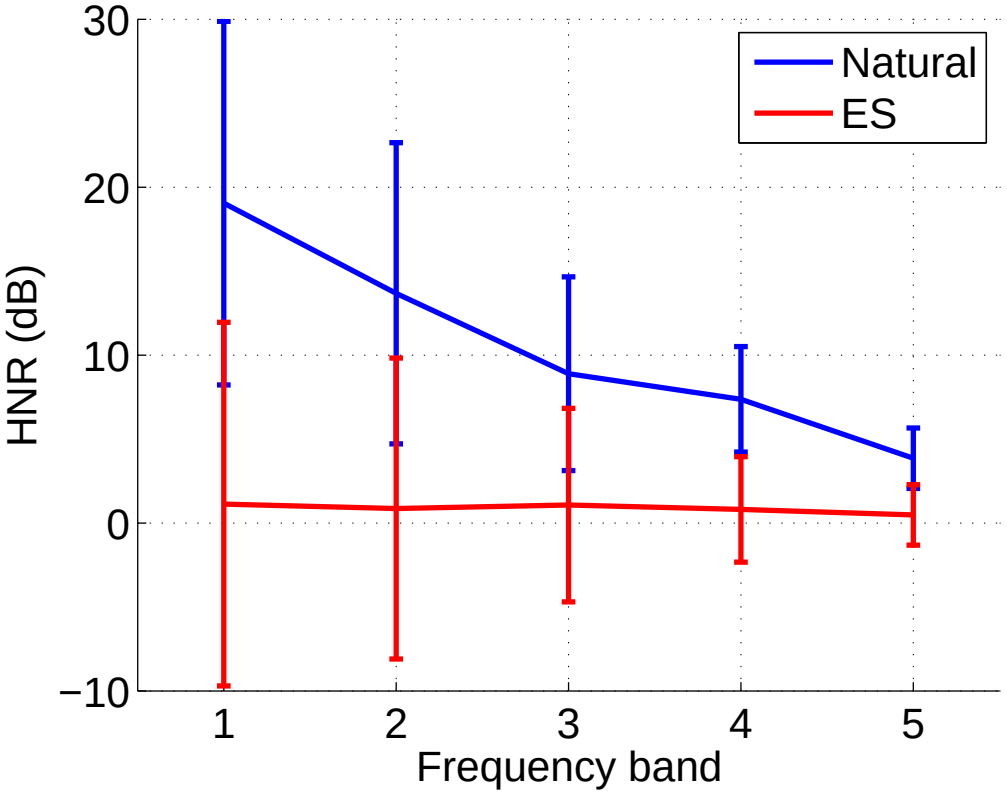

FIGURE 3.12: Harmonic to Noise Ratio (HNR) of natural and ES speech (vowel /a/)

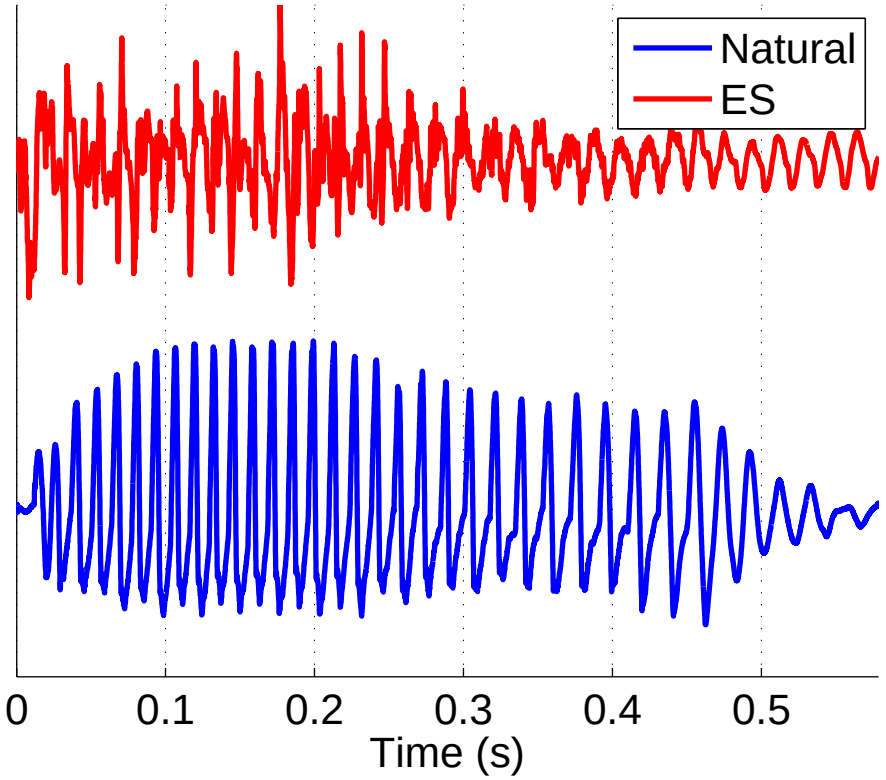

FIGURE 3.13: Source excitation for natural and ES speech (vowel /a/)

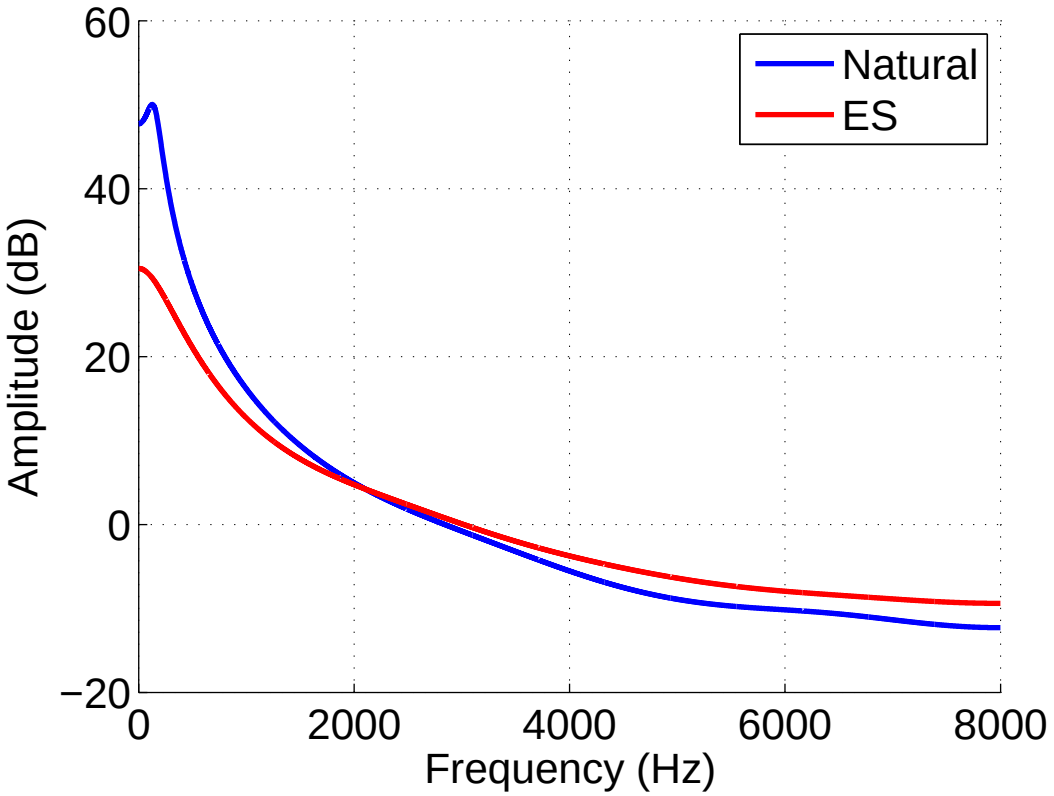

FIGURE 3.14: Source spectrum of natural and ES speech(vowel /a/)

the natural source spectral tilt and has high frequency emphasis, can be observed in Figure 3.14.

Analyzing the vocal tract of both ES and NS, it can be noted that formants for Spanish vowels are shifted upward in the frequency, as well the higher frequency components are emphasized more in ES as shown in the Figures 3.15 3.16 3.17 3.18 3.19.

In short summary, it can be seen from the analysis of both source and vocal tract, the ES has the following problems in comparison to Normal Speech (NS);

- source signal lacks regular fundamental frequency F_0 and source signal resembles whispered speech source signal, i.e. no F_0 or harmonics
- HNR of source signal is not stable because of instability in F_0.
- spectral decay also has high frequency emphasis shape
- formant frequencies are irregularly moved to higher frequencies in spectrum of the vocal tract,
- formant bandwidths also need to be increases as it is reduced in ES,

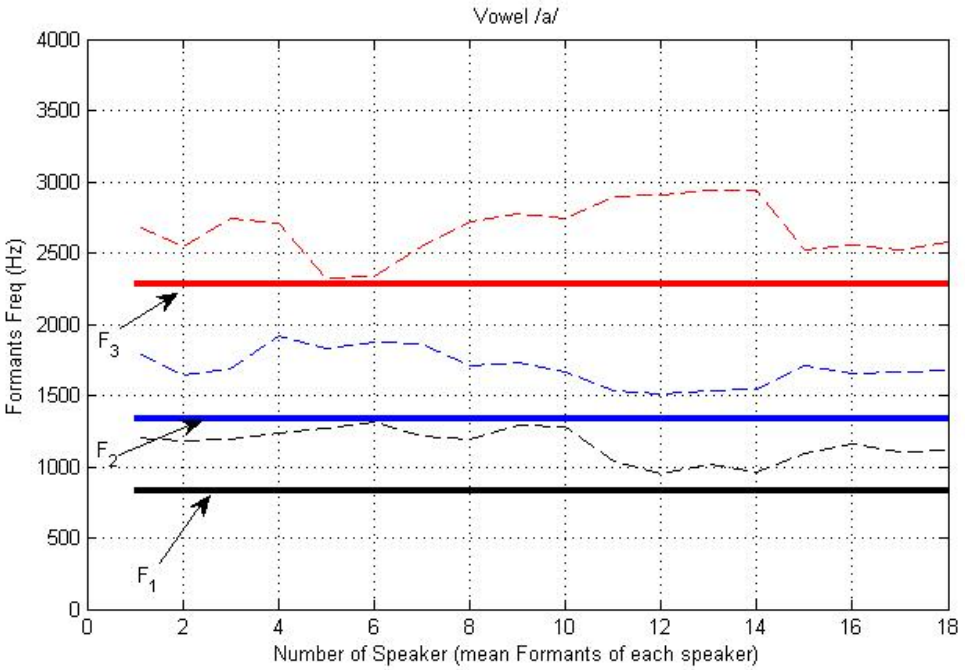

FIGURE 3.15: The formants deviation for the Spanish ES vowel /a/

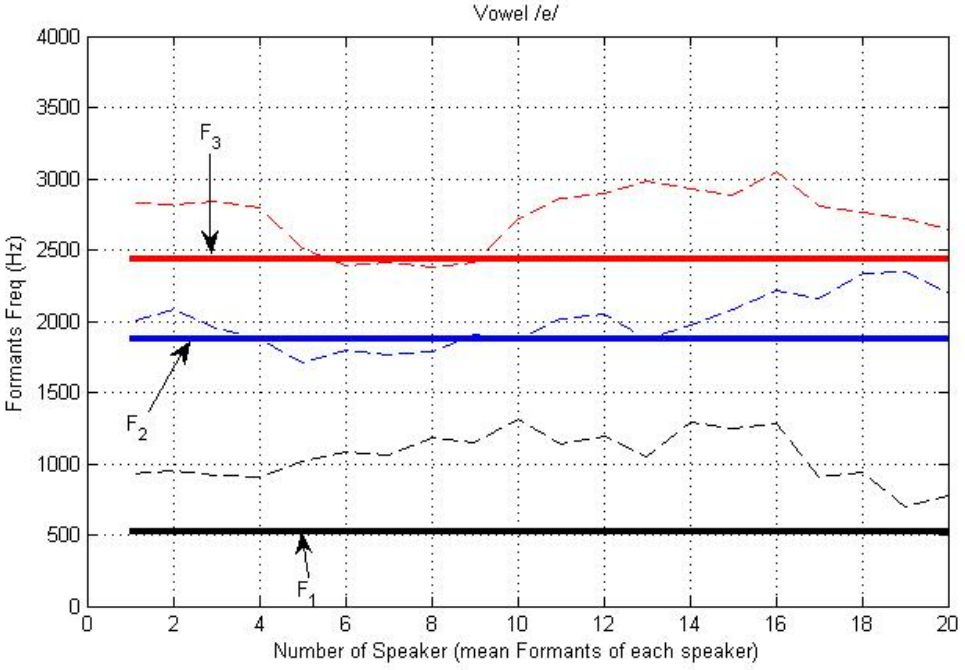

FIGURE 3.16: The formants deviation for the Spanish ES vowel /e/

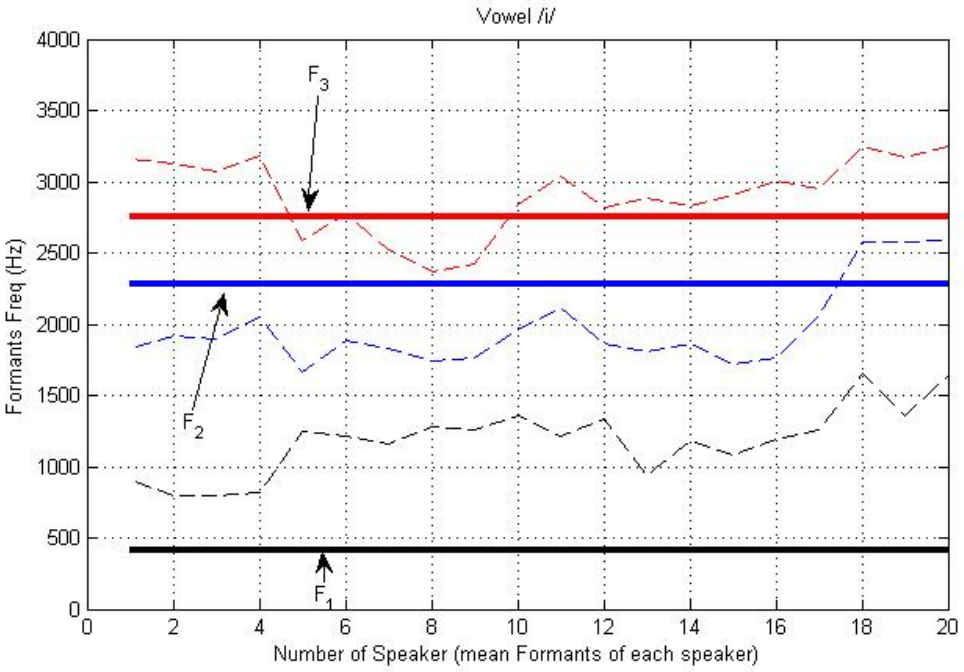

FIGURE 3.17: The formants deviation for the Spanish ES vowel /i/

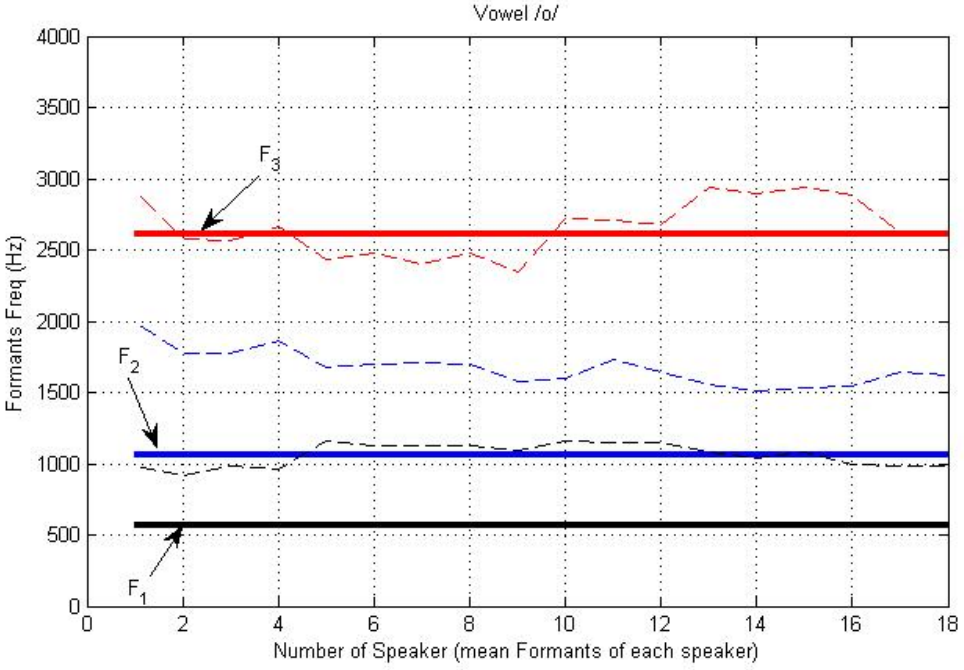

FIGURE 3.18: The formants deviation for the Spanish ES vowel /o/

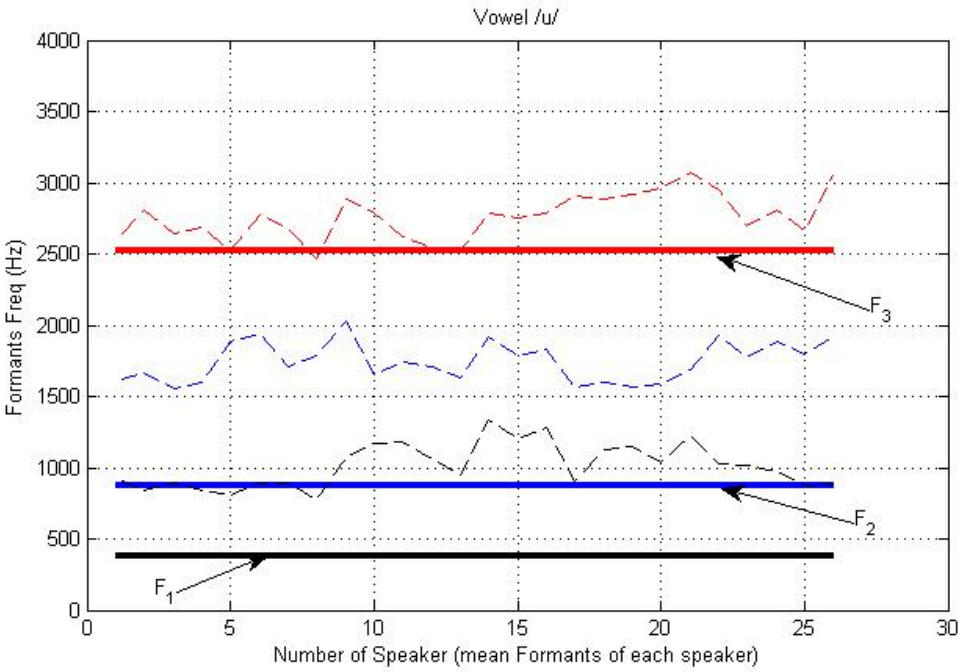

FIGURE 3.19: The formants deviation for the Spanish ES vowel /u/

For better and intelligible ES, it is necessary to transform the source and filter of ES to NS components. This section describes the methods for this transformation both for source and filter components independent of each other.

3.2.1 Source transformation

The source signal in ES is the most effected components, and need to be corrected for the better and intelligible ES. Figure 3.20 shows the steps for the voiced source modification. The natural glottal pulse, interpolated according to the natural fundamental frequency curve extracted from normal speech for glottal pulses for each frame. The gain G of the original ES frame is used to adjust the energy of the source signal. The HNR of any arbitrary normal speech, is used to add noise to avoid robotic like source signal. The spectral tilt is applied to matched the spectrum to desire spectrum. The desired spectrum follows the normal speech spectral tilt for better source signal. The lip radiation and essential part of source signal is applied at the end for the transformed source signal $\hat{g}[n]$. Subsequent sections provide the detailed description of each component of the source signal transformation according to the Figure 3.20.

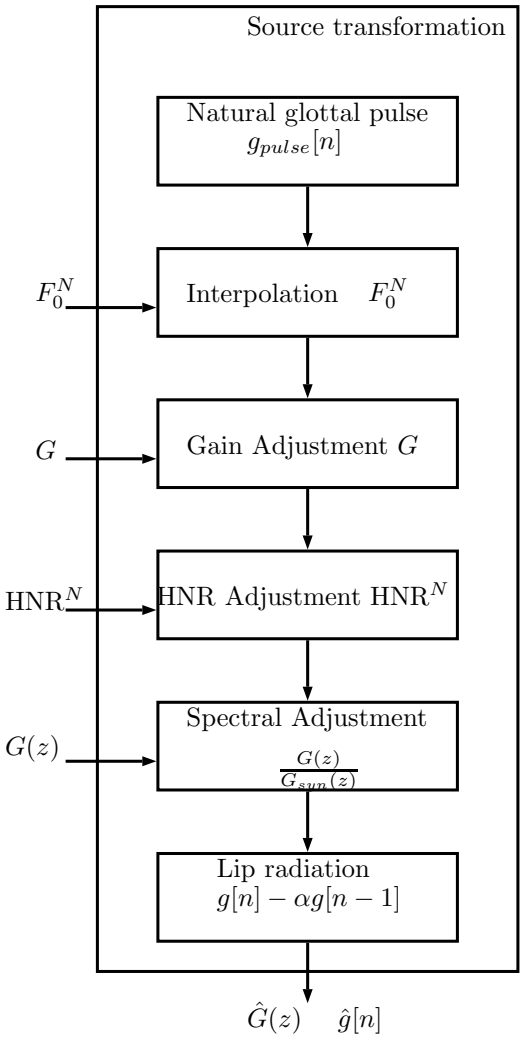

FIGURE 3.20: Source transformation part of the system

3.2.1.1 Natural glottal pulse

The natural glottal pulse extracted from normal speech is used for this purpose and shown in Figure 3.21. This natural glottal pulse also can be model using the 10^{th} order polynomial as:

$$g_{pulse}[n] = \sum_{k=1}^{N} p_k n^{N-k}, \quad N = 11 \text{ and } n = 1 \rightarrow 145 \tag{3.14}$$

where N is polynomial order, n is number of samples for glottal pulse modeling, and p_k are polynomial coefficients and values are: $\{p_1 = -4.18 \times 10^{-18}$, $p_2 = 2.93 \times 10^{-15}$, $p_3 = -8.73 \times 10^{-13}$, $p_4 = 1.42 \times 10^{-10}$, $p_5 = -1.40 \times 10^{-8}$, $p_6 = 8.40 \times 10^{-7}$, $p_7 = -3.02 \times 10^{-5}$, $p_8 = 5.16 \times 10^{-4}$, $p_9 = -0.0064$, $p_{10} = 0.0284$, and $p_{11} = -0.0231\}$.

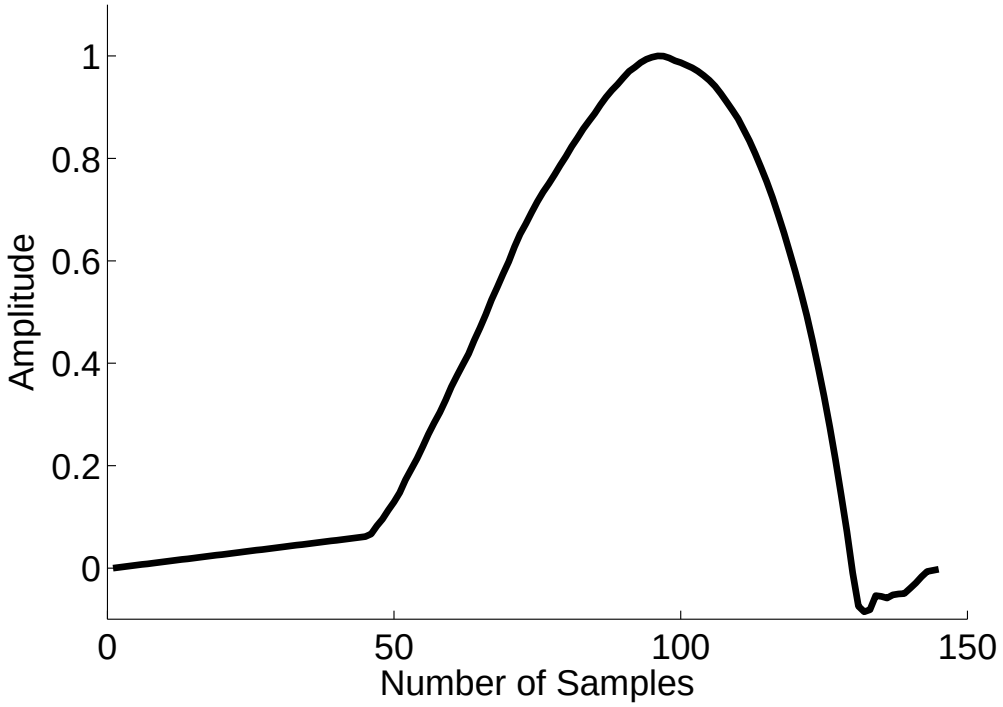

FIGURE 3.21: Natural glottal pulse extracted from normal speech

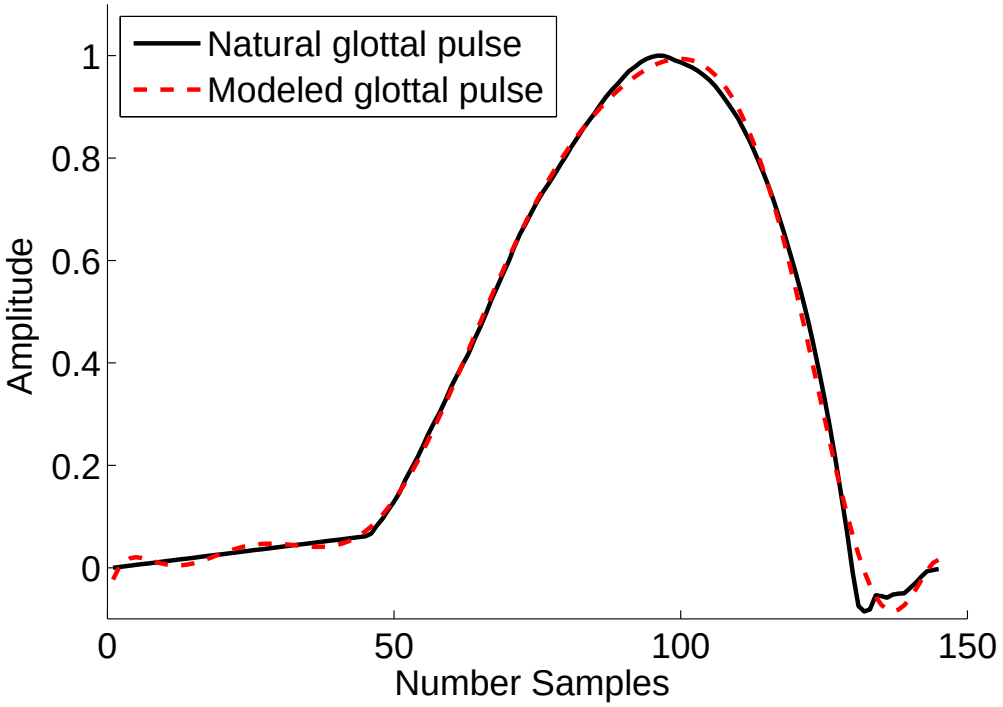

FIGURE 3.22: Natural glottal pulse vs model glottal pulse by 10^{th} order polynomial

3.2.1.2 Interpolation

The F_0 of the ES has irregular shape, and some time it does not exist, so any normal speech F_0 curve can be used. So instead of using the ES F_0, and arbitrary normal speech F_0^N can be used. The natural glottal pulse first interpolated using FFT interpolation method, according to the new F_0^N.

3.2.1.3 Gain adjustment

The F_0^N interpolated pulses source signal $g_{syn}[n]$ is then uses the energy G of the frame for equalizing it to the original frame energy.

$$g_{syn}[n] = Gg_{syn}[n] \tag{3.15}$$

where G is energy of the frame.

3.2.1.4 HNR adjustment

In order to avoid the robotic or machine like source signal, noise is added to the source signal according to the HNR. The estimated HNR of ES is not accurate, any arbitrary normal speech HNR is used in this experiment. The method is first estimate the HNR of $g_{syn}[n]$ as estimated in analysis part of the system. The arbitrary normal speech HNR^N is then used with HNR of $g_{syn}[n]$ using these following steps:

- take FFT of $g_{syn}[n]$
- add random components (white Gaussian noise) according to the HNR^N to the FFT real and imaginary parts.
- take IFFT of the noise modified source signal

In mathematical form it can be:

$$G_{syn}(z) = G_{syn}(z) + Q(z) \tag{3.16}$$

where $G_{syn}(z)$ is the spectrum of $g_{syn}[n]$, and $Q(z)$ is the HNR based noise component.

3.2.1.5 Spectral adjustment

The spectrum of the $g_{syn}[n]$ is almost constant because of synthetic pulses, so it needs to be compensate for target spectrum, and that is $G(z)$. To achieve this, an IIR filter is constructed using [63]:

$$H_m(z) = \frac{G(z)}{G_{syn}(z)} \tag{3.17}$$

where $G(z)$ and $G_{syn}(z)$ are original and synthetic LP source spectrum. The spectrally matched synthetic source signal $g[n]$:

$$g[n] = g_{syn}[n] * h_m[n] \tag{3.18}$$

where $h_m[n]$ is matched filter impulse response, and $*$ is convolution operator. The difference between spectral slopes of natural, ES, and estimated can be observed in Figure 3.23.

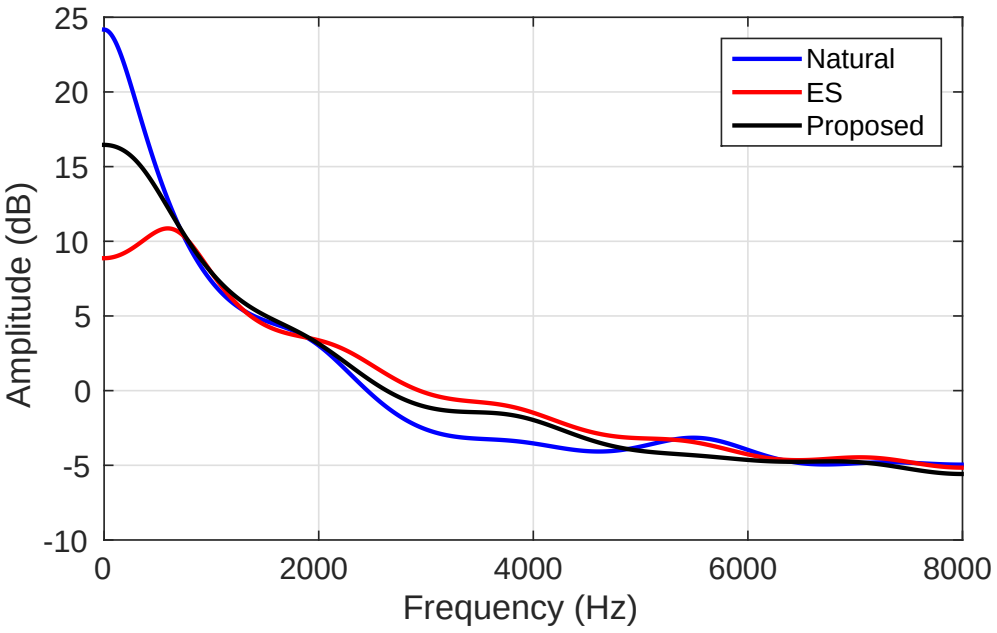

FIGURE 3.23: Spectra of source signal along with natural and ES source signal

3.2.1.6 Lip radiation

Finally lip radiation is applied to get the modified source signal:

$$\hat{g}[n] = g[n] - \alpha g[n-1], \quad 0.96 < \alpha < 1 \tag{3.19}$$

where α is a lip radiation constant. The Figure 3.24 shows the system generated source signal, and it can be seen, that it is closer to natural speech. The spectra of source signal also follows the natural speech spectral tilt as shown in Figure 3.23.

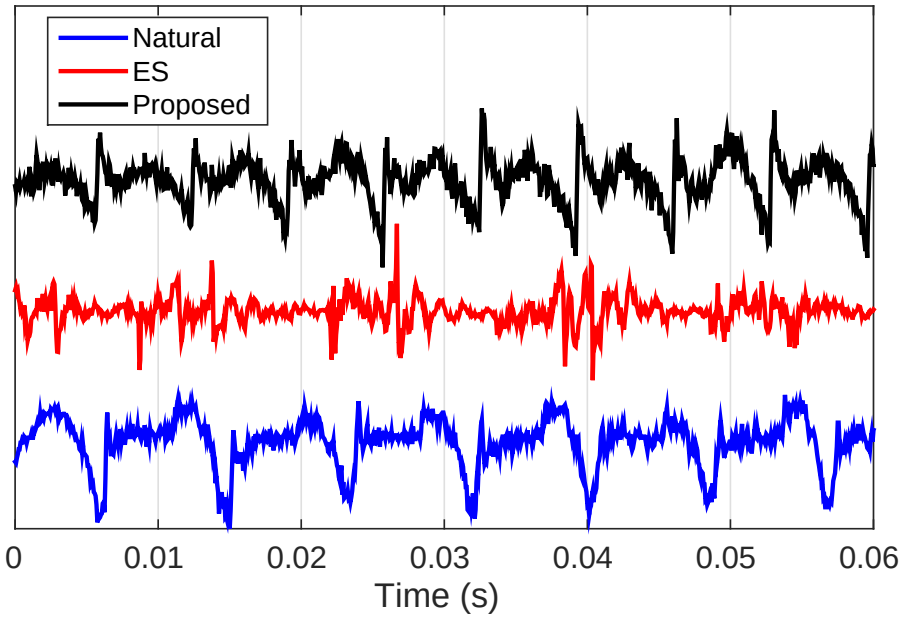

FIGURE 3.24: System generated source signal along with natural and ES source signal

3.2.2 Vocal tract transformation

It can be seen from the vocal tract analysis of ES in comparison to normal speech, it has following problems, which needs to be compensate;

- vocal tract spectra has high frequency emphasis trend
- spectral peaks (formants) of the spectra are moved higher in the frequency in comparison to normal speech
- spectral width (formants bandwidth) of the spectra also decreased, and needed to be increase

The Figure 3.25 shows the proposed algorithm for dealing all the problems mentioned above.

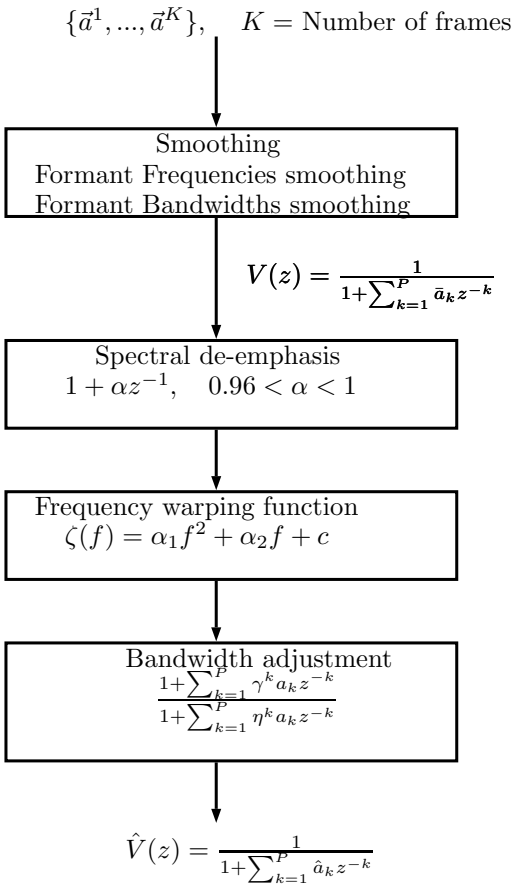

FIGURE 3.25: Vocal tract transformation part of the system

3.2.3 Smoothing

Before applying the different processing to vocal tract, it is compulsory to smooth the formants frequencies and its bandwidth, for smooth spectrum. The median filtering is applied for smoothing purpose, for the curve over all the frames.

3.2.3.1 Spectral de-emphasis

To address the problem of high frequency emphasis, de-emphasis filter is applied;

$$H_{de}(z) = 1 + \alpha z^{-1}, \quad 0.95 < \alpha < 1 \tag{3.20}$$

where α is de-emphasis constant. The enhanced $H_{enh}(z)$ vocal tract transfer function hence is;

$$H_{enh}(z) = V(z)H_{de}(z) \tag{3.21}$$

$$H_{enh}(z) = \frac{1 + \alpha z^{-1}}{1 + \sum_{k=1}^{P} a_k z^{-k}} \tag{3.22}$$

where $V(z)$ is the vocal tract transfer function, and a_k are the LP coefficients of order P.

3.2.3.2 Frequency warping function

In order to solve the problem of shifted spectral peaks (formants) in higher frequency, a compensation function is needed, which moves the peaks to lower frequency according to normal speech. For this purpose an second order Frequency Warping Function (FWF) is calculated;

$$\zeta(f) = \alpha_1 f^2 + \alpha_2 f + c \tag{3.23}$$

where $\alpha_1 = 6.079 \times 10^{-5}$, $\alpha_2 = 0.5553$, and $c = 60.280$. The transformed frequency $\hat{f}$ using FWF is given as:

$$\hat{f} = \beta\zeta(f), \quad \beta = 1, f = 0 \rightarrow \frac{f_s}{2} \tag{3.24}$$

where β is a constant, f is original frequency and $\hat{f}$ is transformed frequency, and f_s is a sampling frequency. The first four formants of Spanish ES and normal speech vowels /a/, /e/, /i/, /o/ and /u/ are compared and difference between these is modeled by FWF curve as shown in Figure 3.26. The Figure 3.27 shows the original and frequency

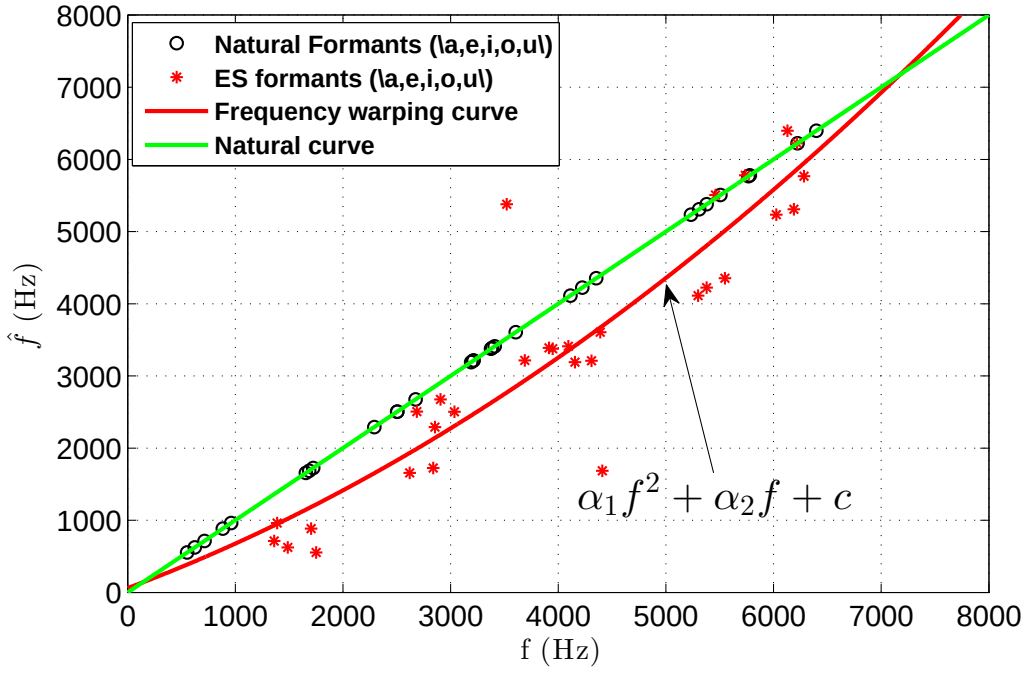

FIGURE 3.26: Frequency Warping Function (FWF).

warped spectra.

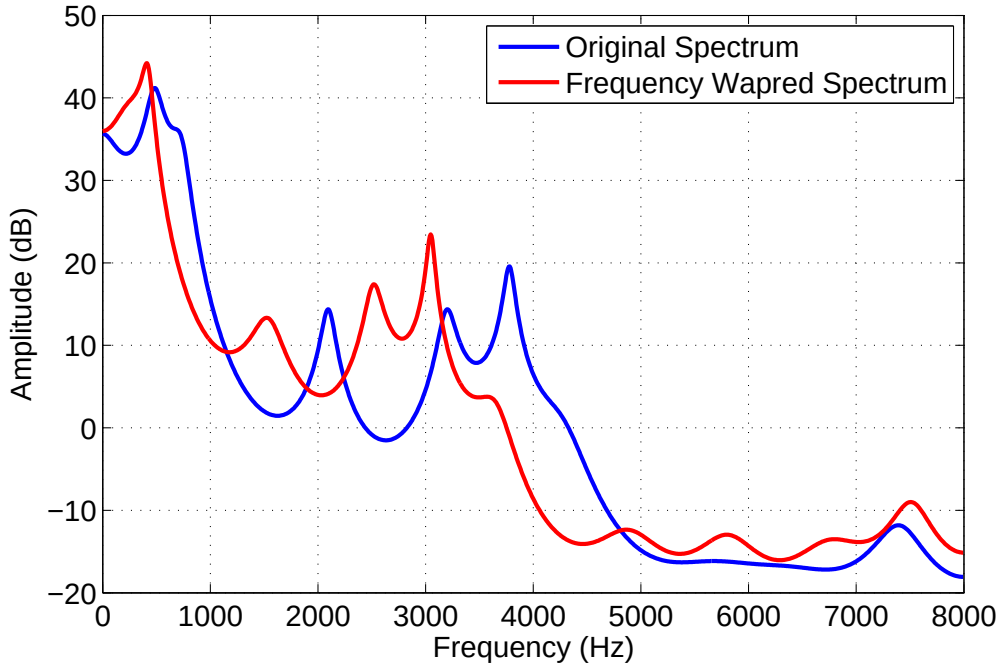

FIGURE 3.27: Frequency warped spectra.

3.2.3.3 Bandwidth adjustment

To increase or decrease the spectral peaks width (formants bandwidth), an IIR transfer function is used [172]:

$$H_{bw} = \frac{1 + \sum_{k=1}^{P} \gamma^k a_k z^{-k}}{1 + \sum_{k=1}^{P} \eta^k a_k z^{-k}} \tag{3.25}$$

where γ and η are spectral bandwidth controlling constants. The spectral bandwidth increases for $\gamma > \eta$, otherwise it decreases. The Figure 3.28 is showing the effect of varying the values of γ and η, and showing how it effects the spectral peaks (bandwidth), and can be seen that when the value of γ is greater than η such as ($\gamma = 0.99$ and $\eta = 0.98, 0.96$) the bandwidths of peaks reduce, and in opposite such as ($\gamma = 0.96, 0.98$ and $\eta = 0.99$) bandwidth of peaks increase.

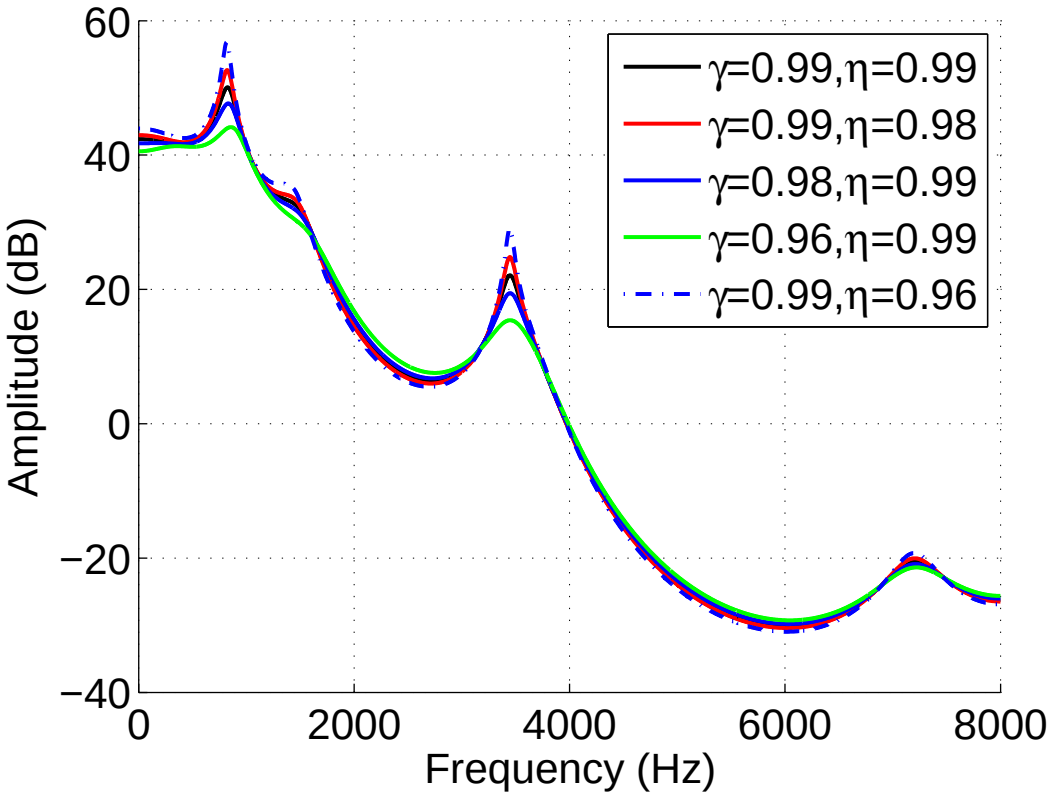

FIGURE 3.28: Spectral bandwidth modification.

3.3 Synthesis

Voiced speech

The modified source signal $\hat{g}[n]$ and vocal tract impulse response $\hat{v}[n]$ are then convolved for enhanced version of ES $\hat{s}[n]$ for voiced speech;

$$\hat{x}[n] = \hat{g}[n] * \hat{v}[n] \tag{3.26}$$

$$\hat{X}(z) = \hat{G}(z)\hat{V}(z) \tag{3.27}$$

Unvoiced speech

For the unvoiced speech, the unvoiced vocal tract, and white noise $\varepsilon[n]$ are convolved according to the gain of the frame G;

$$\hat{x}[n] = G\varepsilon[n] * v[n] \tag{3.28}$$

The Figure 3.29 shows the steps for synthesizing the speech frames.

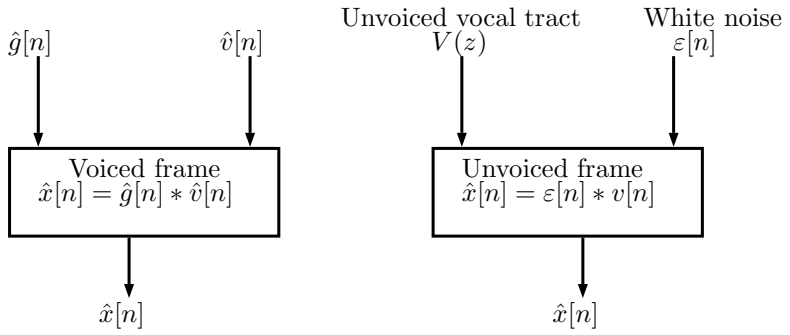

FIGURE 3.29: Synthesis part of the system

The overall effect of the system to the vocal tract filter, and source signal can be seen in the Figures 3.30, and 3.31 respectively. Figure 3.30 shows that the spectral peaks and spectral tilts follows the normal speech spectral peaks and tilts. By observing the source signal from Figure 3.31, proposed system follows the normal speech source signal.

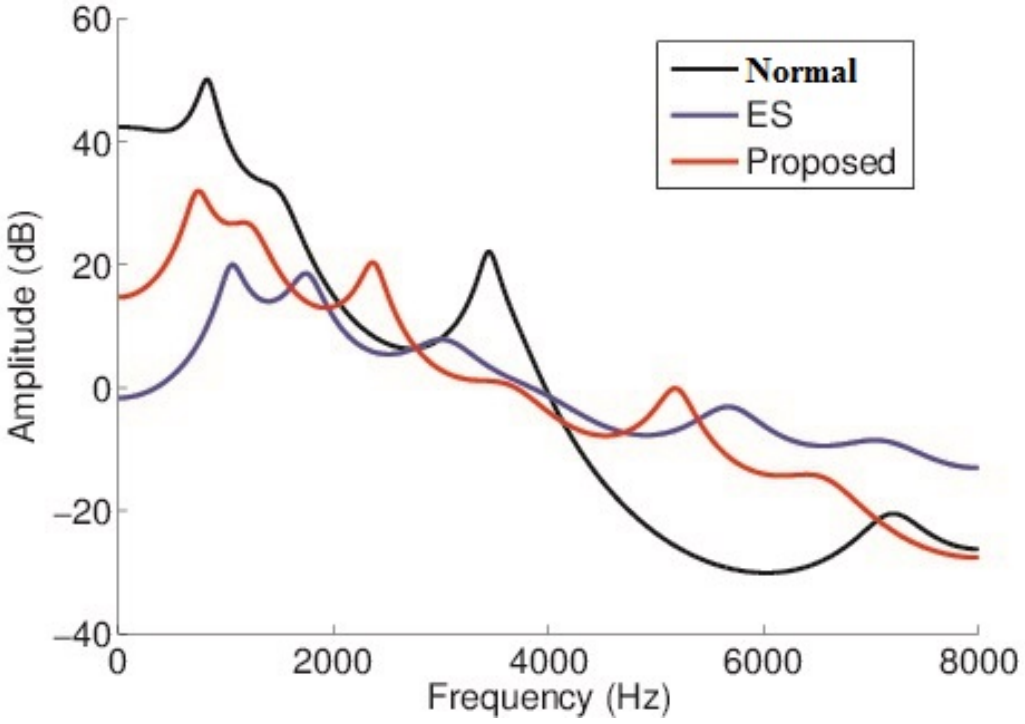

FIGURE 3.30: Linear prediction spectra of normal speech, ES and enhanced with system

The system transforms the source and vocal tract filter optimally, closer to the normal speech source and vocal tract filter. The effect of the system can be clearly seen from the spectrogram of the speech signal. Figures 3.32, and 3.33 are shown the spectrogram of uprocessed and processed with system for the vowel /a/.

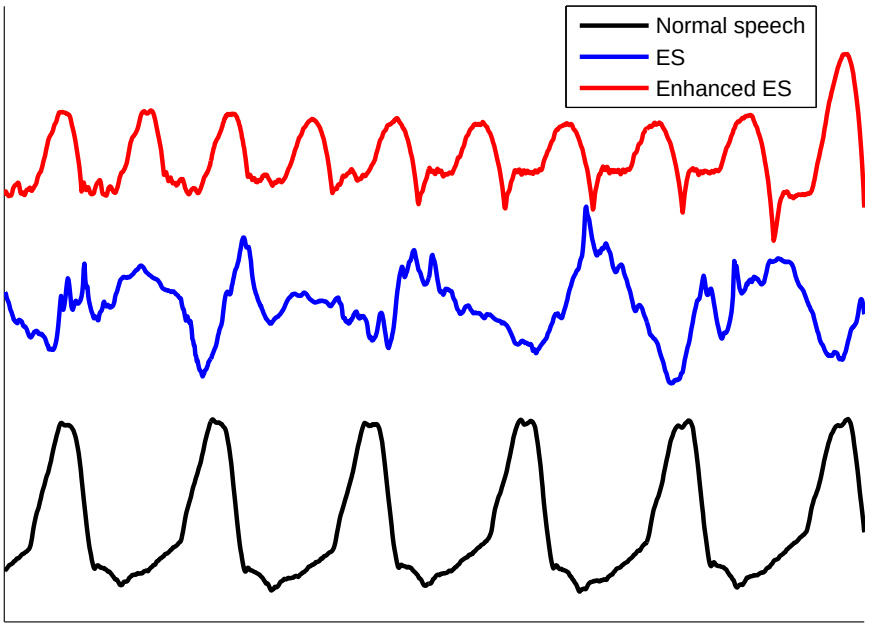

FIGURE 3.31: Source signal of normal speech, ES, and enhanced ES

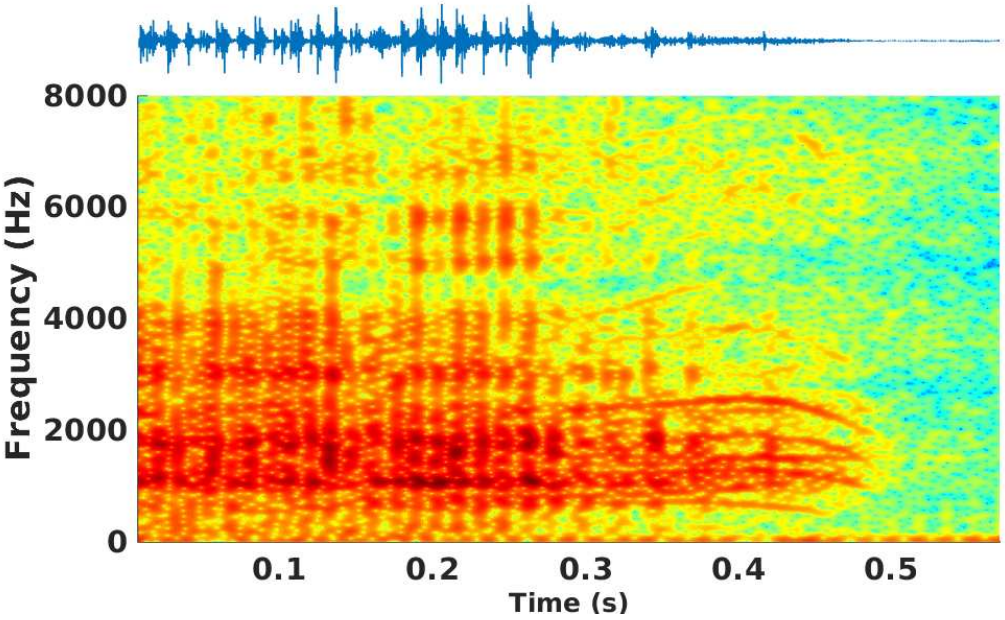

FIGURE 3.32: Spectrogram of unprocessed vowel /a/

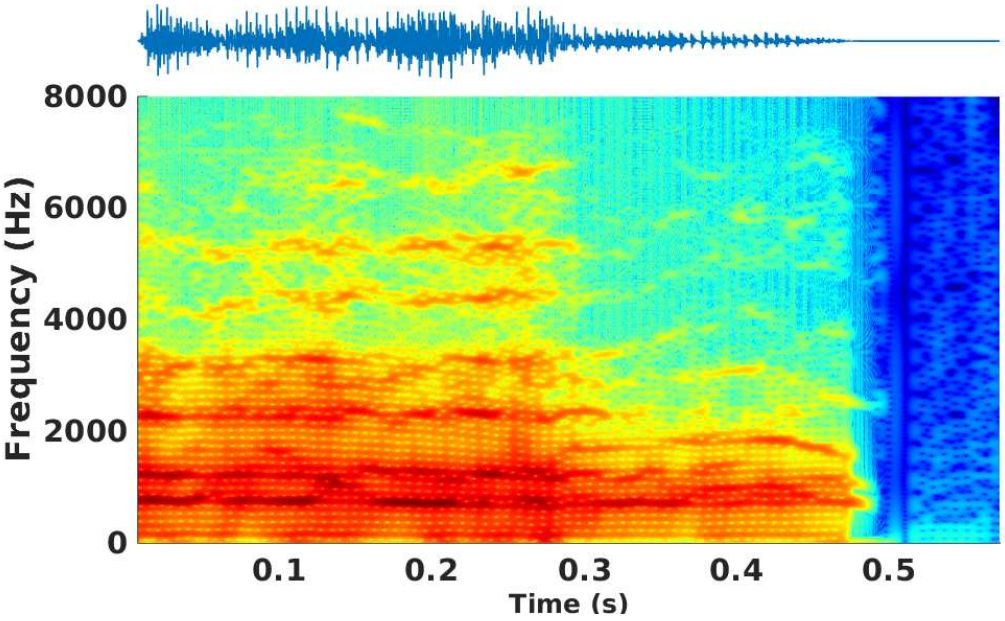

FIGURE 3.33: Spectrogram of vowel /a/ processed with the proposed system

3.4 Chapter Summary

This chapter has addressed the problems of ES in a sense of source and filter components, i.e. what source and vocal tract filter are missing in ES. A new method for source and vocal tract filter transformation to normal speech source and vocal tract filter has been presented. The source which lacks periodicity, and corresponds to whispered speech source, uses natural glottal pulse extracted from normal speech, with the normal speech fundamental frequency and HNR. The original frame energy and source spectrum has been used for transforming the source to normal speech source. The vocal tract filter has been transformed to normal speech vocal tract by addressing the problems of higher frequencies emphasis, spectral peaks shifting and spectral peaks width reduction. These problems has been solved by spectral de-emphasizing, frequency warping function and spectral peaks bandwidth enlargement.

Chapter 4

Results

4.1 Speech Database

In order to assess the system ability to enhance the ES intelligibility, large number of ES words were used. The words list contains 28 Spanish two-syllable (CVCV) ES words and sustained vowes /a/, /e/, /i/ , /o/, and /u/. The words list reflects the frequently used syllabic structure of the Spanish language. The 28 words, and vowels are divided into two different parts;

- pino, cita, tira, liso, rima, milla, dique, letra, vega, seda, templo, perla, cero, petaca, musa, nube, poda, zona, rosa, goma, bodega, ganso, fase, gasa, jaspe, papa, mama, chino
- /a/, /e/, /i/, /o/, /u/ - sustained vowels

There were six male speakers, with an average age of 55. Three of the speakers are using ES restoration method for the last six years, and serving as an instructor to other people, while two of other speakers using ES for the last three years, and one speaker usage time is only nine months. Each speaker uttered the word 3 times. The database contains a total of 504 utterances of words, and 90 utterances of vowels. In total the database has 594 words and vowels.

The speech samples were recorded in the "Asociación Vizcaína de Laringectomizados y Mutilados de la Voz" in Bilbao (Figure 4.1). The association provides facilities to the largyngectomees in restoring their speech, and help psychologically to be in the normal life. The speech were recorded individually in an audiomatric booth with a high quality microphone, Cardiod Condenser microphone AT200 from audio-technica [173] (Figure 4.2). The mouth to microphone distance was 15 cm. The speech samples were

FIGURE 4.1: Asociación Vizcaína de Laringectomizados y Mutilados de la Voz

recorded with sampling frequency of 44.1 kHz, which were down-sampled to 16 kHz for computational efficiency. The speakers were given the list of words on the paper, and instructed to read the words with the pause of approximately four seconds between the words. The words were then separated by freely available audio editing software WavePad [174]. The database is limited to male speakers, because the association does not have any female speaker.

FIGURE 4.2: A high quality Cardiod Condenser microphone At2020 from audio-technica (adapted from [173])

4.2 Experiments

The proposed system was tested with two subjective listening tests and one with objective parameter. The subjective listening tests were based on Mean Opinion Score (MOS), and preference test, while the HNR was used as objective evaluation. The visual presentation of the results has been seen using the Spectrogram. The speech database was divided into vowels and words group. The words are grouped into 4 random group of each with 7 words. The groups were made for avoiding too much words for one listeners. There were 50 listeners, for each group 10 listeners participated. The age of the listeners were 20 to 32, with an average age of 28. The listeners have no knowledge of speech signal processing so the results are unbiased. The tests were conducted in a room which has no or low level background noise. As there are lot of utterances for each speech sample, so to avoid boredom and fatigue among the listeners words were divided into 4 groups and vowels;

- /a/, /e/, /i/, /o/, /u/ [vowels]
- pino, cita, tira, liso, rima, milla, dique [group-1];
- letra, seda, vega, templo, perla, cero, petaca [group-2];
- musa, nube, poda, zona, rosa, goma, bodega [group-3]
- ganso, fase, jaspe, papa, mama, chino, gasa [group-4];

The MOS is a widely used perceptual quality test for speech signals for the long time. The listener sit in the quite room and score the perceived speech signal according to ITU-T recommendation P.80 [175]. According to [175] the listener rate the heard speech using the scale 1 to 5. The 1 being worst and 5 as best. The table 4.1 shows the rating scheme and corresponding speech quality. In the first listening test, listeners were presented

MOS	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

TABLE 4.1: Mean Opinion Score (MOS).

with original and processed speech samples randomly without knowing which one is the

original and which one is the processed. The task of the listeners were to rate the speech samples according to the table 4.1.

The second listening test was preference test based on selecting preferred speech sample among the given utterances. The listeners were asked to selected one of the speech samples among the original and processed samples based on which one they would rather to listen. Based on the number of answers to this preference, results were presented in percentage. To avoid the biasing of the results, listeners were not aware of which one is processed or which one is the original speech sample. The proposed system presented in this thesis was compared with the reference system [83] for the quality assessment purpose. The reference system [83] is described in the next section briefly in order to understand the differences between the proposed and reference system, i.e. how source and vocal tract filter are decomposed and modified. The detailed description of the reference system, one should see the [83].

4.2.1 Reference system

The Figure 4.3 shows the reference system [83] used for comparison purpose in this thesis. The reference system first decomposes the speech signals into frames. Each frame is classified as voiced and unvoiced frame using zero crossing rate, frame energy and the Bindex. The unvoiced frames are not altered or modified, and used at the synthesizing step without modifying. The voiced frames are decomposed into source and vocal tract filter components using linear prediction analysis [14]. The source and vocal tract filter components are processed separately. The source signal is used to calculate fundamental frequency F_0 using the Esophageal Voice-Modified Auto-Correlation Method (EV-MACM) (is the modified version of Modified Auto-correlation Method (MACM))[83]. The F_0 curve is then smoothed, and used with LF source model [25, 176] for source synthesis, i.e. for generating the modified source signal for the synthesizing. The vocal tract filter formants are extracted using the roots extraction of the prediction coefficients of the vocal tract [42]. The bandwidths of the formants are calculated from the roots of the prediction coefficients of the vocal tract [42]. The formants bandwidths are enlarge using the perceptual weighting filter [172]. The spluttering noise from the higher frequencies of the vocal tract is also reduced. Once the vocal tract filter and source signal are modified, they are synthesized using the linear prediction synthesizer for the enhanced version of the ES [14]. The modified vocal tract and source are then synthesized using linear prediction synthesizer [14]. The Figure 4.3 shows the system components for both source and vocal tract filter. The reference system is adapted according to the proposed system. Both the reference and proposed system were configured with the similar parameters, such as, frame size, frame overlap, and prediction

order p etc. The reference system although originally only assessed perceptually using the subjective listening test of Mean Opinion Score (MOS).

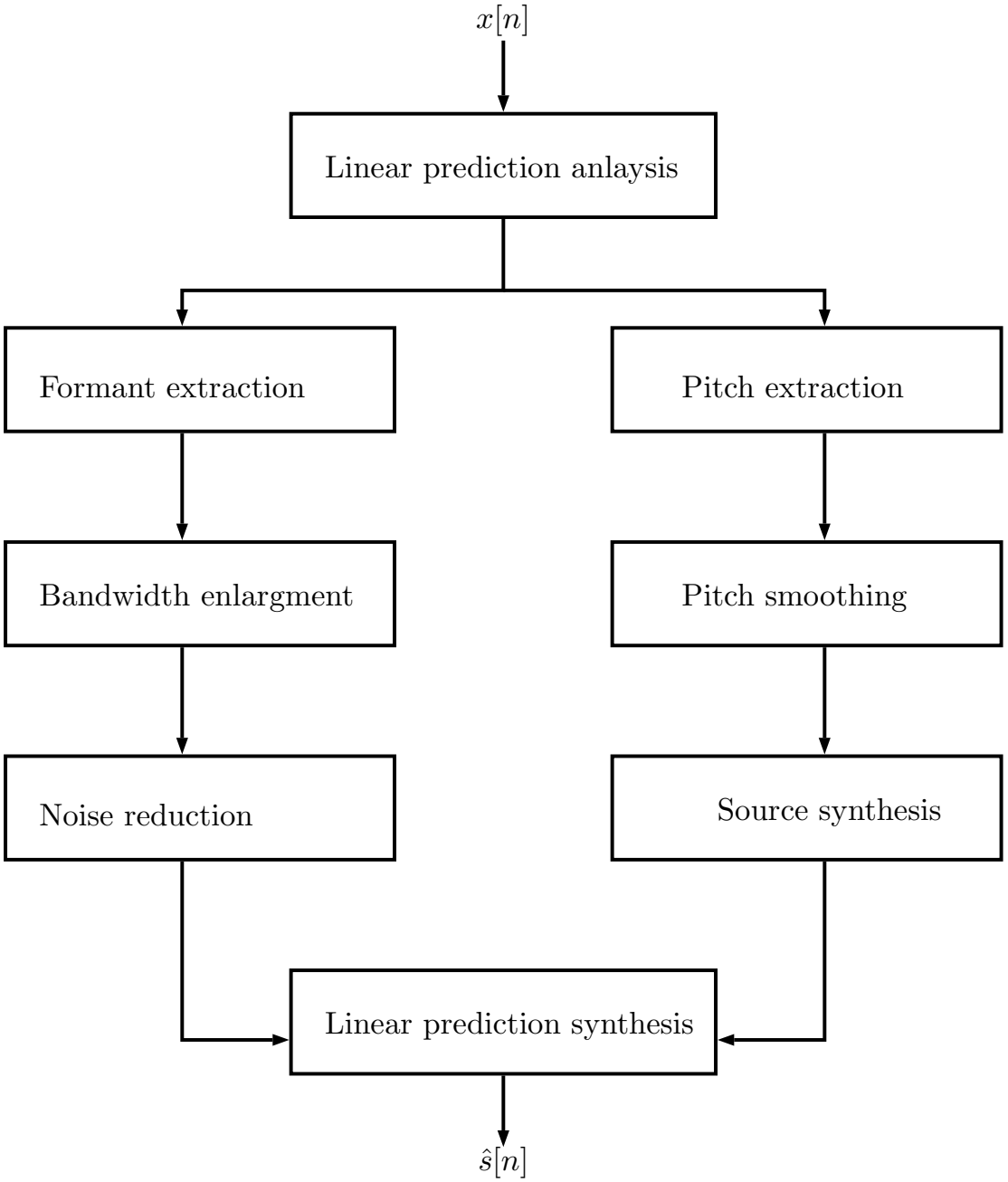

FIGURE 4.3: Reference system (adapted from [83])

4.2.2 System Configurations

In order to ensure comparability of the proposed and reference systems, identical configurations were used for the system tunning. The common configurations such as, the frame size was set to 30-msec, the frame shift was 5-msec. The vocal tract filter was

modeled with the prediction order p of 30. The source spectrum $G(z)$ was used only in proposed system, and its prediction order was set to 10.

4.3 Subjective listening test

The proposed system performance was evaluated by two subjective listening test. The first listening test was a quality evaluation test based on Mean Opinion Score (MOS) and second the preference test.

4.3.1 Mean Opinion Score (MOS)

The system in first phase only test with Spanish ES vowels /a/, /e/, /i/, /o/ and /u/. The listener were given the random list of vowels. To make the test unbiased, listeners were not told, about the processed and original utterances. The average values of the vowels across all the listeners were calculated for MOS, and reported. Figure 4.4 shows the average MOS for all the vowels, where it can be seen that the proposed system outperform the reference system significantly. The mean MOS for the proposed system falls between 2.8 to 3.5, which is a good score for ES.

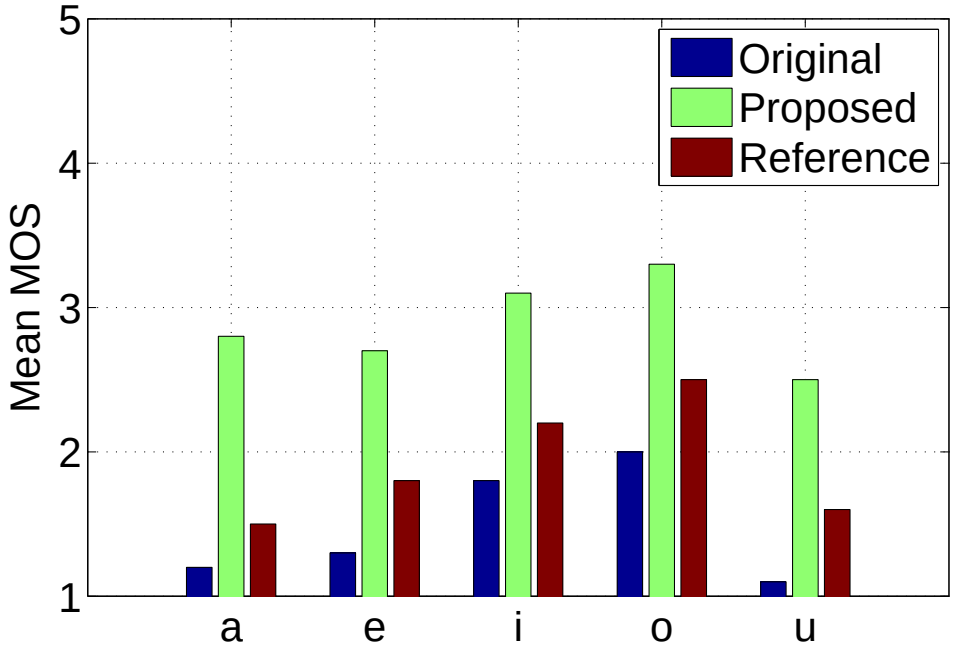

FIGURE 4.4: *Results of the MOS test for all the vowels.*

The MOS of the normal vowels and the ES vowels processed by the proposed and reference system were statistically evaluated using p-values for significant correlation.

The significance correlation p-values were calculated using the Matlab statistical toolbox. The small values of p, say less than 0.05, indicates that the correlation between samples is significant. The table 4.2 shows the p-values for the significance correlation among the processed and normal speech vowels. It is noticeable from the table that proposed system has significance correlation with normal speech as the p-value is less than 0.05. From the table, it is also observed that vowel /o/ also has the p-value less than 0.05 for the reference system. But the overall the proposed system p-values shows that it has much closer relation with normal speech.

-	Vowel	p-value	Significance (p<0.05)
Original	/a/	0.4511	x
Reference system	/a/	0.0613	x
Proposed system	/a/	0.0033	√
Original	/e/	0.3072	x
Reference system	/e/	0.0823	x
Proposed system	/e/	0.0135	√
Original	/i/	0.4461	x
Reference system	/i/	0.1429	x
Proposed system	/i/	0.0061	√
Original	/o/	0.0531	x
Reference system	/o/	0.0421	√
Proposed system	/o/	0.0051	√
Original	/u/	0.8123	x
Reference system	/u/	0.0753	x
Proposed system	/u/	0.0312	√

TABLE 4.2: Comparison of original and processed ES vowels with normal speech vowels

In the second phase of perceptual evaluation using MOS, the list of words from the database were used. The task of the listeners was to rate the words using MOS scale. The words list was divided into four groups to ease the listening test. The words of every group were heard by 10 listeners. The table 4.3 shows the average MOS for all the words. It is noticeable from the table that the proposed system has the average MOS above the 3 most of the time in comparison to reference system. The proposed system therefore, outperform reference system significantly.

The average MOS of the different groups is also calculated and shown in the Figure 4.5. It is shown from the Figure 4.5, that the MOS among the group as well has the high score as compare to reference system. The average MOS among the group is more the 3 for the proposed system and for the reference system speech samples and original samples it is 2.5 and 1.5 respectively, which implies the low quality of the ES. Therefore,

Word	Original	Reference system	Proposed system	Group
Pino	1.8	2.3	3.6	1
Cita	1.9	2.8	3.1	1
Tira	1.7	3.1	2.8	1
Liso	2.1	2.9	3.3	1
Rima	2.3	2.7	3.4	1
Milla	2.1	2.3	2.9	1
Dique	1.8	2.3	3.1	1
Letra	2.3	2.4	3.1	2
Vega	1.8	2.6	2.2	2
Seda	1.1	2.3	3.4	2
Templo	1.5	2.9	3.6	2
Perla	1.9	2.9	3.1	2
Cero	1.4	2.1	2.5	2
Petaca	1.1	1.2	2.3	2
Musa	1.3	3.1	3.6	3
Poda	2.1	3.2	3.3	3
Zona	1.8	2.9	3.1	3
Rosa	2.3	2.1	3.7	3
Goma	2.7	2.6	3.9	3
Bodega	2.4	2.8	3.1	3
Nube	2.5	2.6	3.5	3
Ganso	1.9	2.6	3.1	4
Fase	1.6	3.1	3.4	4
Jaspe	2.6	2.9	2.8	4
Papa	2.9	2.5	3.1	4
Mama	2.1	2.4	3.4	4
china	1.6	1.8	2.7	4
Gasa	2.1	2.9	3.7	4

TABLE 4.3: Mean Opinion Score (MOS).

the subjective listening test using the MOS scaling has outperform the reference system significantly.

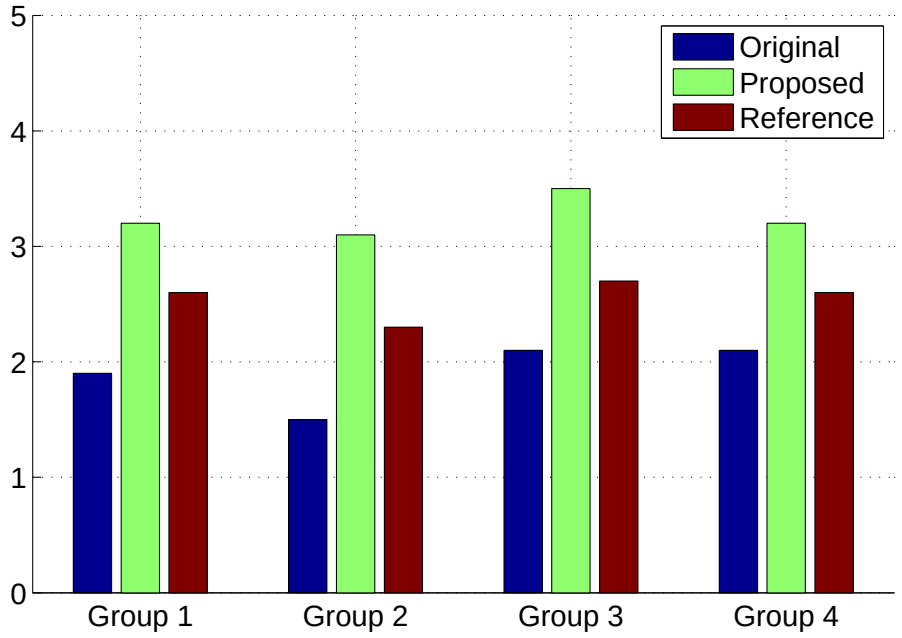

FIGURE 4.5: *Average MOS for the groups for words list.*

Summary

The perceptual listening test conducted for the proposed system and measured using MOS, has shown that the proposed system MOS is fall between 3 to 4 for all the vowels and the words. The higher value of MOS shows that the proposed system has significantly improved the intelligibility of the ES speech samples, and according to MOS table 4.1 the processed speech samples shows that the quality is good. As the words list from the database was evaluated in groups to avoid the fatigue among the listeners, it is observed as well that the proposed system has also improved the MOS among groups, and the MOS falls in the value of good scale range.

4.3.2 Preference Test

The preference test is another type of perceptual evaluation test. In this test, listeners task was to select the sound which they prefer to listen among the different utterance. The preference test also was conduct in groups to reduce the length of the words for one listener. For each group there were 10 listeners, and were given the original, and processed words with proposed and reference system without knowing which one is the original and which one is the processed one. In first phase of the test only vowels were evaluated. Based on the preference test on vowels, it can be seen that the proposed

system has the highest preference for vowels collectively, indicated by the middle bar in the Figure 4.6. The proposed system for all the vowels has shown 76% preference in comparison to reference system which has only 20% preference. The Figure 4.6 shows the data of the preference test by combining all the vowels. In the second phase of

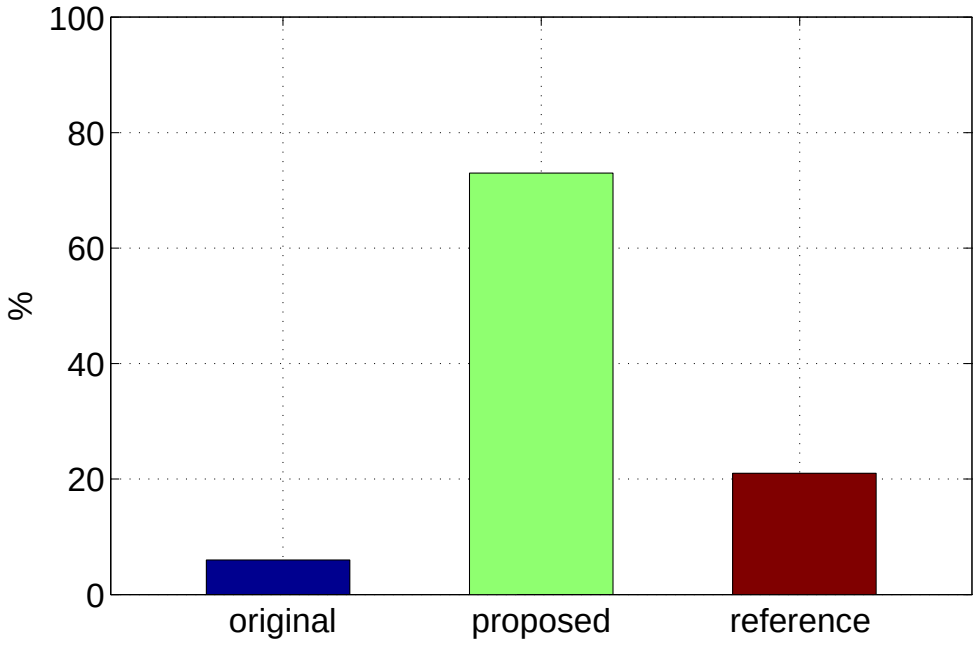

FIGURE 4.6: *Results of the preference test.*

the preference test, words were used. The results of the preference test for second phase are presented in table 4.4. The table indicates that the proposed system has the higher preference in comparison to reference and original speech samples. From the Figure 4.7, the results are presented for all the groups. The average percentage of all the groups shows significant results for the proposed system, as the preference for the proposed system most of the time is above the 50%, in comparison to reference and original samples which has preference always less than 30%. Therefore even among the groups, the results have the similar trend as the individual speech sample, and it has been indicated clearly in table 4.4 and the Figure 4.7.

Summary

The proposed system assessment using the second listening test of preference, where listeners were asked, to select which utterance of vowels and words they prefer to listen. The results on vowels has shown that proposed system processed speech samples were selected most of the time. The results on words list also shows that the proposed system processed speech samples were preferred more than 50%.

Word	Original (%)	Reference system (%)	Proposed system (%)	Group
Pino	17.20	22.70	58.10	1
Cita	19	28	53	1
Tira	17	31	52	1
Liso	21	29	50	1
Rima	18	27	55	1
Milla	14	32	54	1
Dique	22	39	39	1
Letra	13	24	63	2
Vega	18	26	56	2
Seda	11	23	66	2
Templo	15	29	56	2
Perla	19	29	52	2
Cero	14	21	65	2
Petaca	11	12	77	2
Musa	13	34	53	3
Poda	22	29	49	3
Zona	17	35	48	3
Rosa	11	21	68	3
Goma	8	21	71	3
Bodega	12	32	56	3
Nube	25	29	46	3
Ganso	15	20	65	4
Fase	11	21	68	4
Jaspe	19	20	61	4
Papa	14	63	23	4
Mama	21	25	55	4
Chino	28	23	49	4
Gasa	18	26	56	4

TABLE 4.4: Preference score in percentage.

4.4 Objective Evaluation

The proposed system was objectively assessed with Harmonic to Noise Ratio (HNR). The HNR shows the periodicity in the speech signal. The higher the value of HNR, the higher the signal has periodic components [113]. A freely available Matlab based software called VoiceSauce [160] was used for calculating HNR in different bands. As the most of the periodic components of speech signals falls in lower frequencies, the HNR in first band (0-2000 Hz) was used for HNR results. The HNR for unvoiced frame was set to zero, and the mean of HNR from all the frames was taken, and reported.

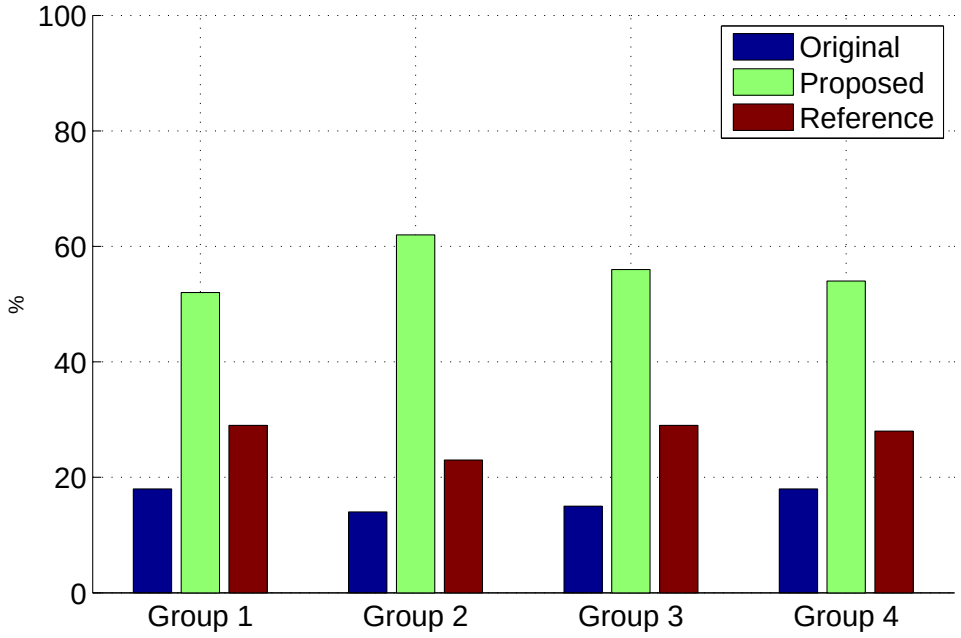

FIGURE 4.7: *Results of the average preference test score for all the groups.*

4.4.1 Harmonic to Noise Ratio (HNR)

In the first phase of the objective assessment of the proposed system, vowels HNR were calculated. Figure 4.8 shows the mean HNR for all the vowels. It is indicated from the figure that proposed system has improved the HNR significantly.

In the second phase of the HNR calculation, the HNR was calculated for the words list from the database. The HNR is calculated by setting the HNR of unvoiced part of the speech to zero, and mean value of voiced parts of speech HNR is taken. The table 4.5 shows the mean HNR for all the words using the VoiceSauce [160]. It is shown from that table that the HNR for the words has significant improvement for the proposed system over the reference and original speech samples.

Summary

The proposed system objectively evaluated using the HNR, an objective measure to measure periodicity in the speech samples. The results for both vowels and for the words from the database has shown that the proposed system improved the HNR significantly in comparison to reference and original speech samples.

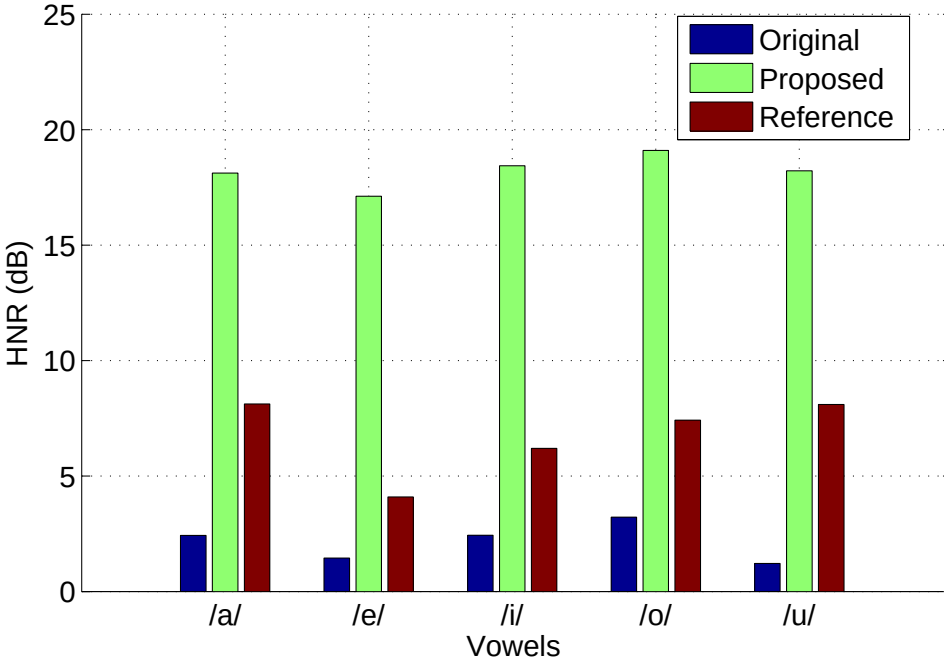

FIGURE 4.8: Mean Harmonic to Noise Ratio (HNR) for all the vowels

Word	Original (dB)	Reference system (dB)	Proposed system (dB)
Pino	2.12	6.25	18.12
Cita	1.45	7.23	17.23
Tira	3.45	8.23	19.11
Liso	-2.43	2.31	18.33
Rima	1.56	5.8	18.45
Milla	5.89	10.23	17.98
Dique	-4.89	5.23	18.24
Letra	-2.45	-0.03	19.01
Vega	-0.416	1.24	18.12
Seda	5.42	12.48	17.22
Templo	4.23	8.92	18.27
Perla	-6.21	5.89	16.89
Cero	6.81	16.33	19.89
Petaca	2.44	15.78	16.23
Musa	-2.34	10.11	19.87
Poda	5.21	10.24	20.11
Zona	1.73	13.53	19.23
Rosa	1.13	12.13	18.45
Goma	-0.08	2.12	15.41
Bodega	1.24	13.42	16.89
Nube	-2.58	15.82	21.22
Ganso	1.45	16.23	18.12
Fase	-5.12	8.72	17.23
Jaspe	1.19	10.78	18.98
Papa	8.41	10.23	19.34
Mama	-2.01	2.51	20.12
Chino	-5.28	12.34	19.31
Gasa	1.28	13.61	17.23

Table 4.5: Mean Harmonic to Noise Ratio (HNR).

4.5 Spectrogram

In order to assess the quality and enhancement in speech samples, visual inspection of the speech samples give good ideas about the movement of vocal tract and source signal. The spectrogram provides the speech signal as a time frequency grid. The spectrogram shows the variations of speech signal in frequency over the time scale. Therefore, it is used in this thesis for visualizing the speech enhancement.

The spectrogram were used for visual analysis of speech signal, where the modification to source and filter components can be seen easily, i.e. Figures 4.9, 4.10, 4.11 show the spectrograms of unprocessed vowel /a/, and processed by proposed and reference systems respectively. From the spectrograms, it can be seen clearly that the proposed system moves and smooth the vocal tract curve, as well the source periodicity as well. Beside that, proposed system also reduces noise in higher frequency, which is shown by blue in the spectrograms.

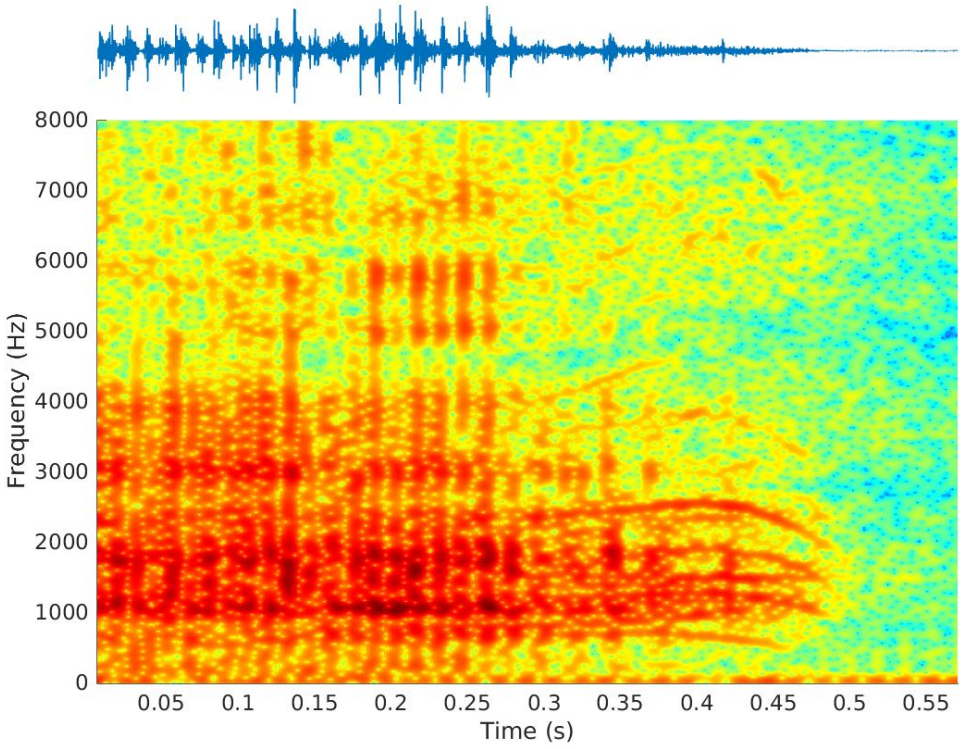

FIGURE 4.9: Spectrogram of the unprocessed vowel /a/

The spectrograms for the unprocessed vowel /e/ is shown in Figure 4.12, and the corresponding spectrograms of the vowel, processed with proposed and reference systems are shown in Figures 4.13, 4.14.

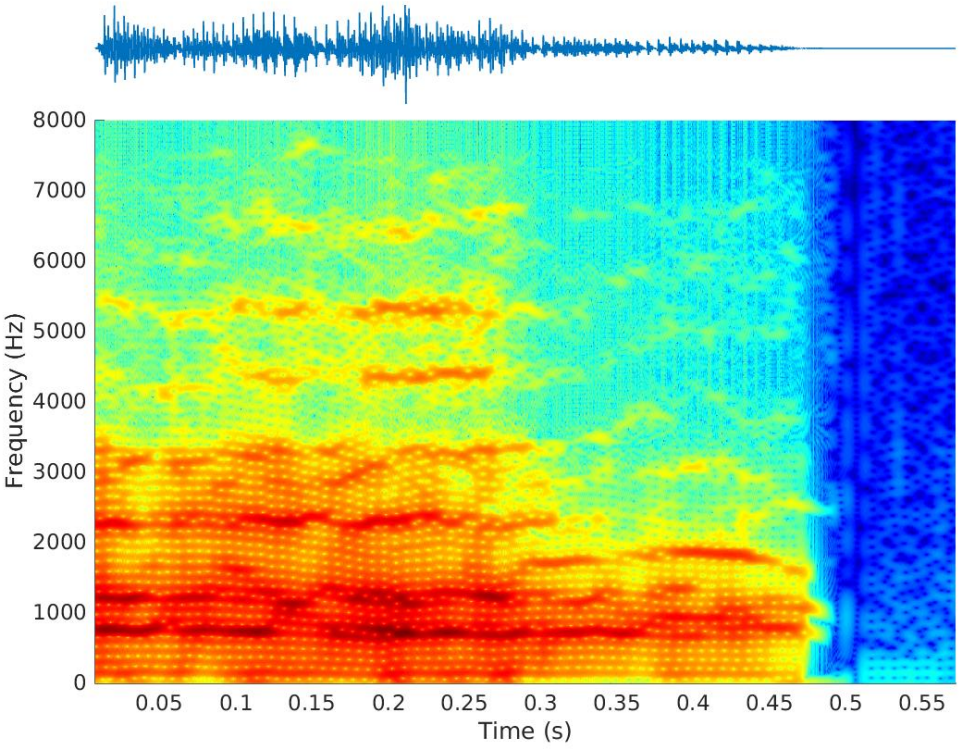

FIGURE 4.10: Spectrogram of the vowel /a/ processed with the proposed system

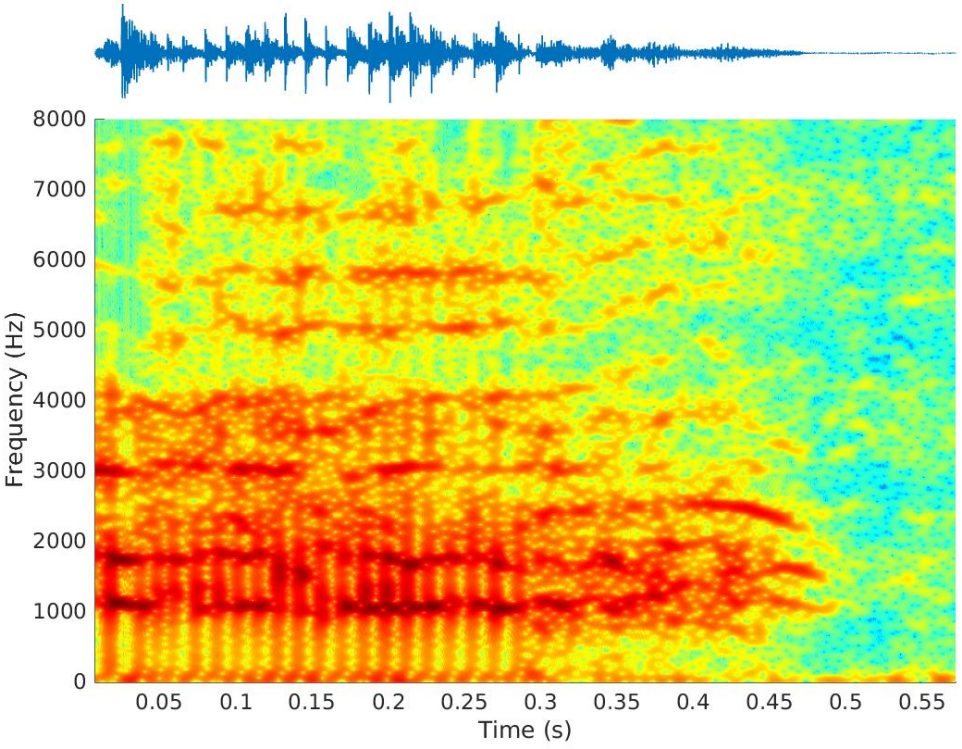

FIGURE 4.11: Spectrogram of the vowel /a/ processed with the reference system [83]

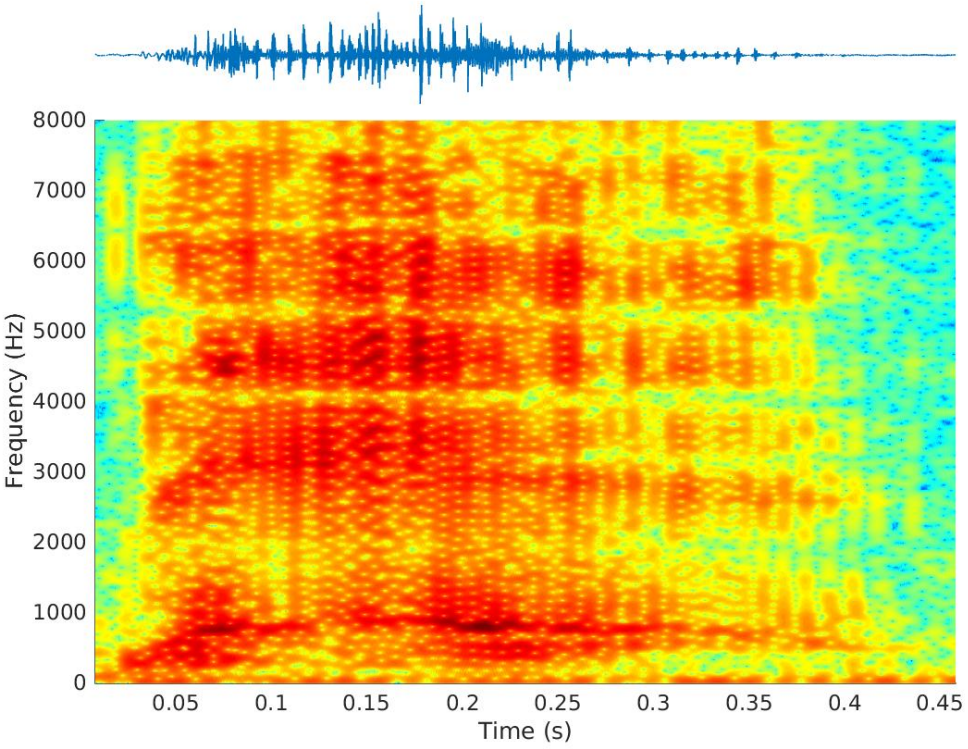

FIGURE 4.12: Spectrogram of the unprocessed vowel /e/

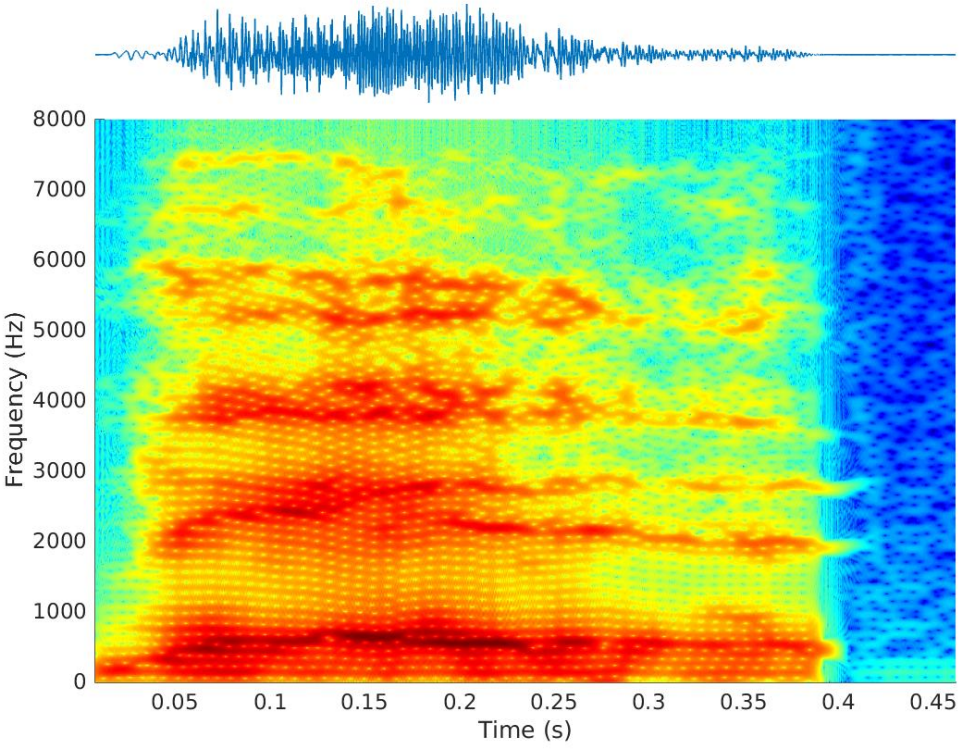

FIGURE 4.13: Spectrogram of the vowel /e/ processed with the proposed system

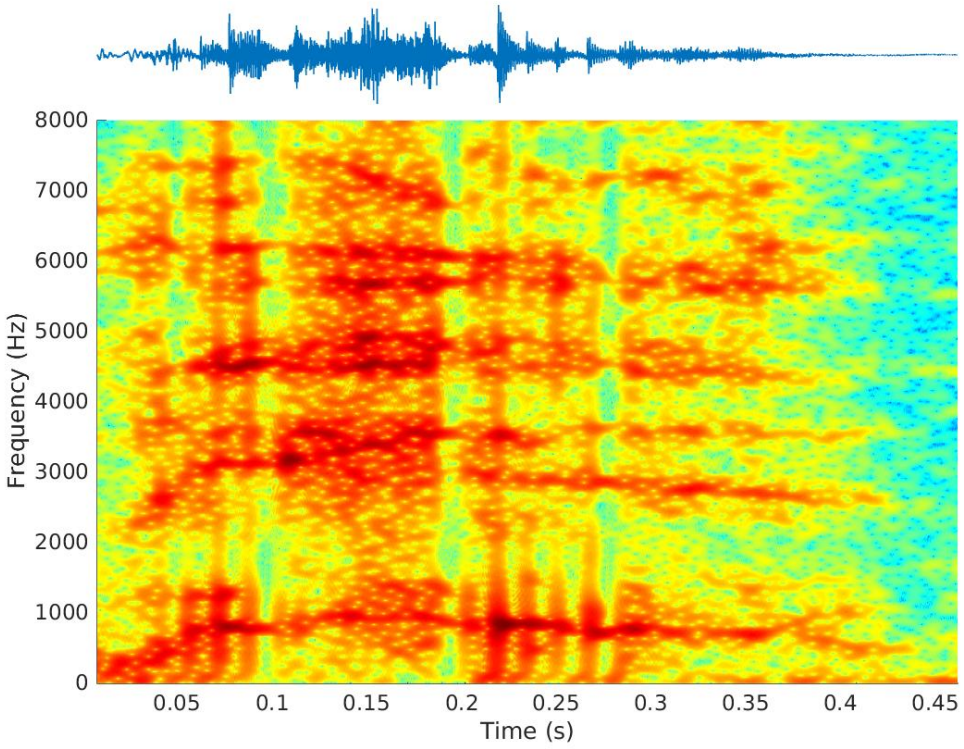

FIGURE 4.14: Spectrogram of the vowel /e/ processed with the reference system [83]

For other vowels /i/, /o/, and /u/ the spectrograms are shown in Figures 4.15 to 4.23.

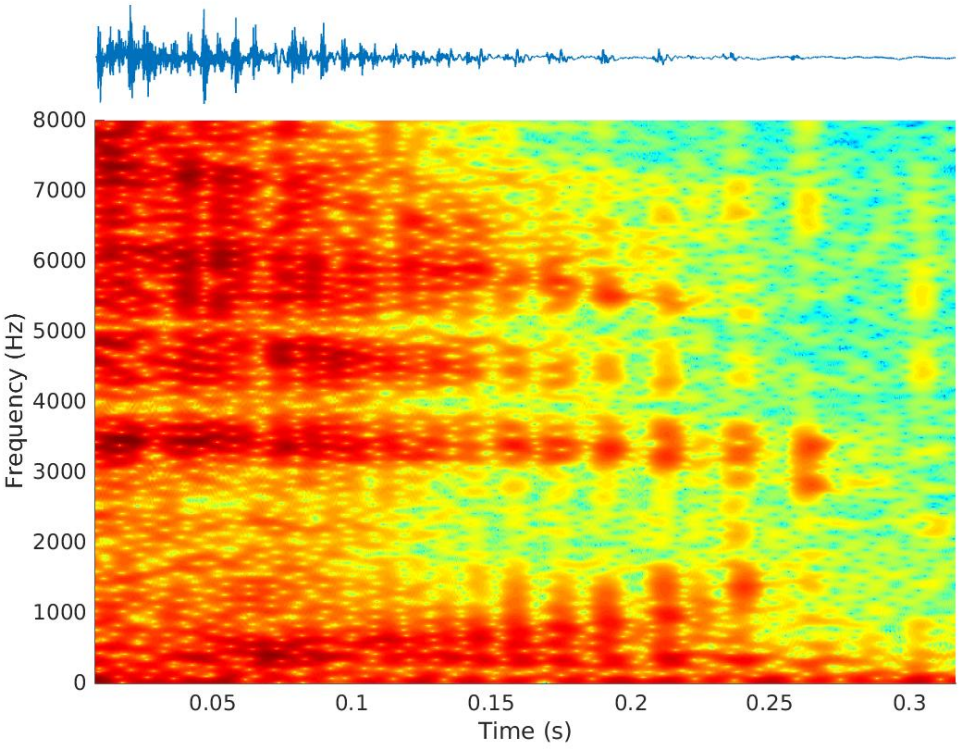

FIGURE 4.15: Spectrogram of the unprocessed vowel /i/

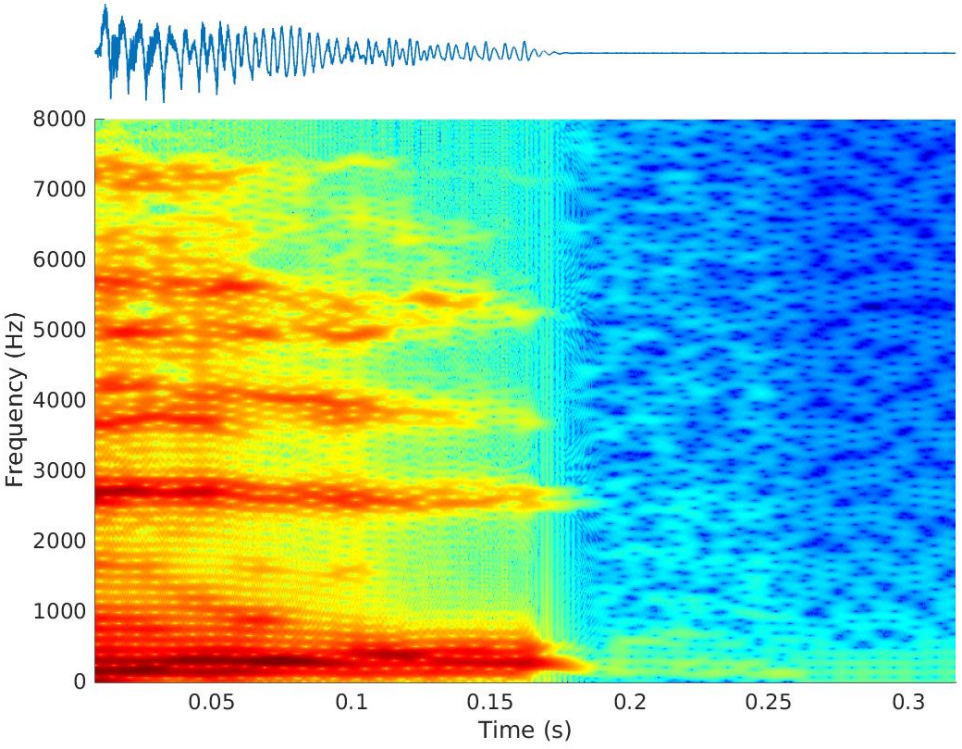

FIGURE 4.16: Spectrogram of the vowel /i/ processed with the proposed system

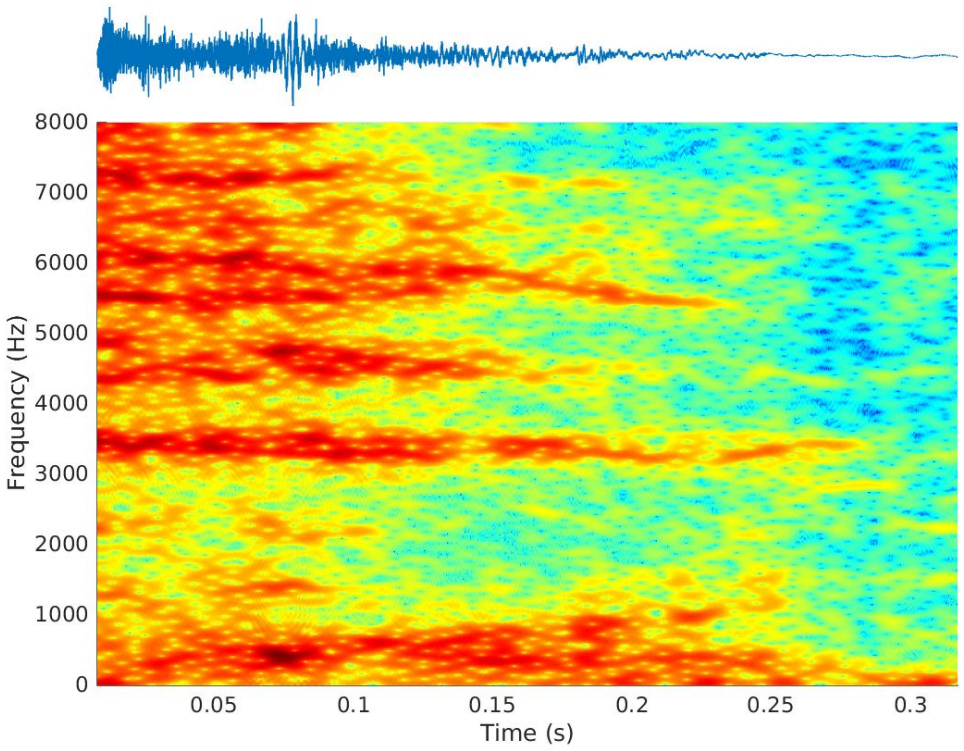

FIGURE 4.17: Spectrogram of the vowel /i/ processed with the reference system [83]

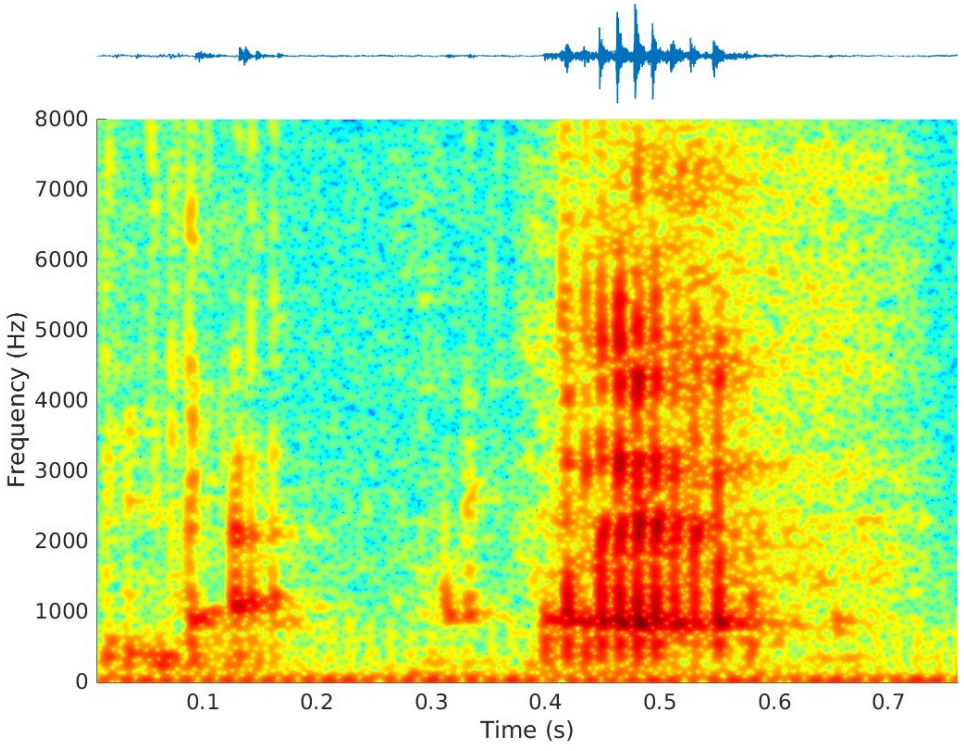

FIGURE 4.18: Spectrogram of the unprocessed vowel /o/

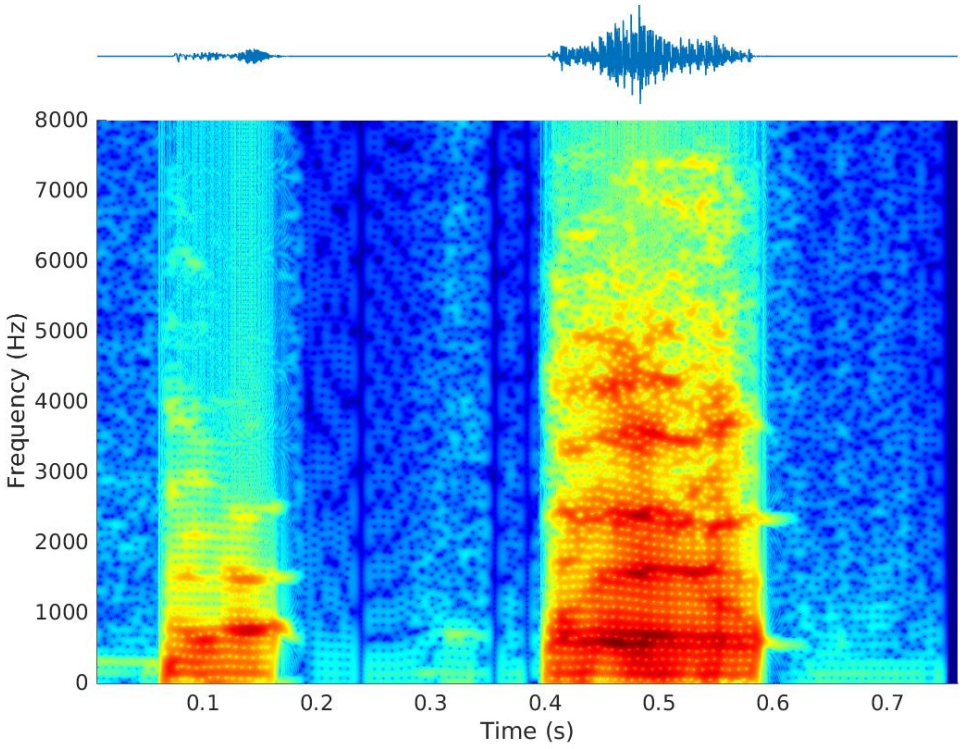

FIGURE 4.19: Spectrogram of the vowel /o/ processed with the proposed system

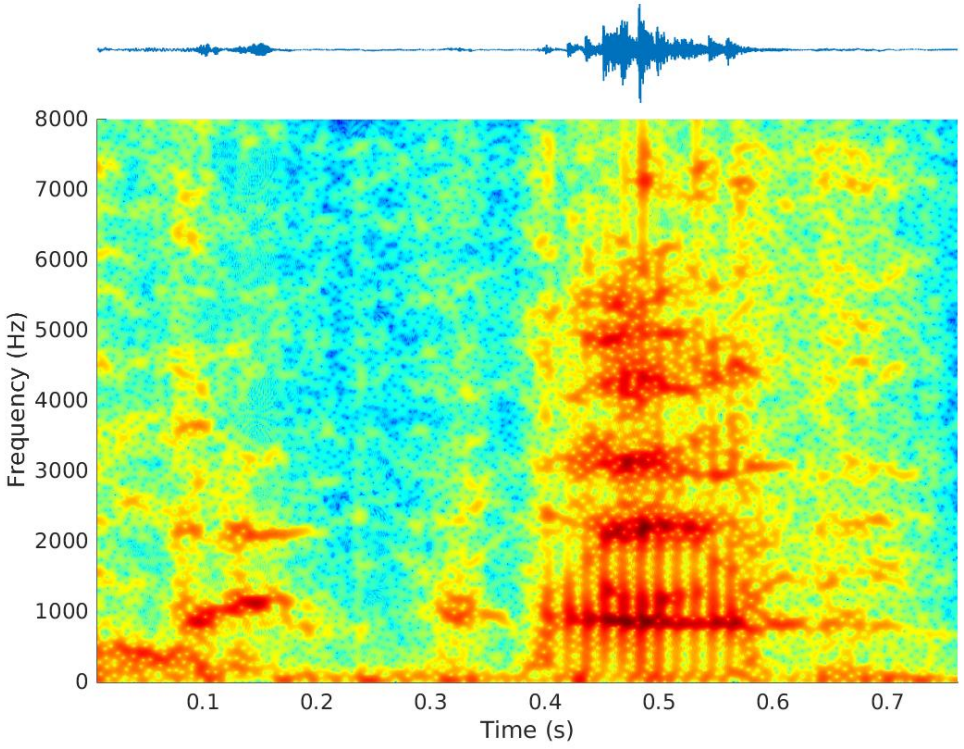

FIGURE 4.20: Spectrogram of the vowel /o/ processed with the reference system [83]

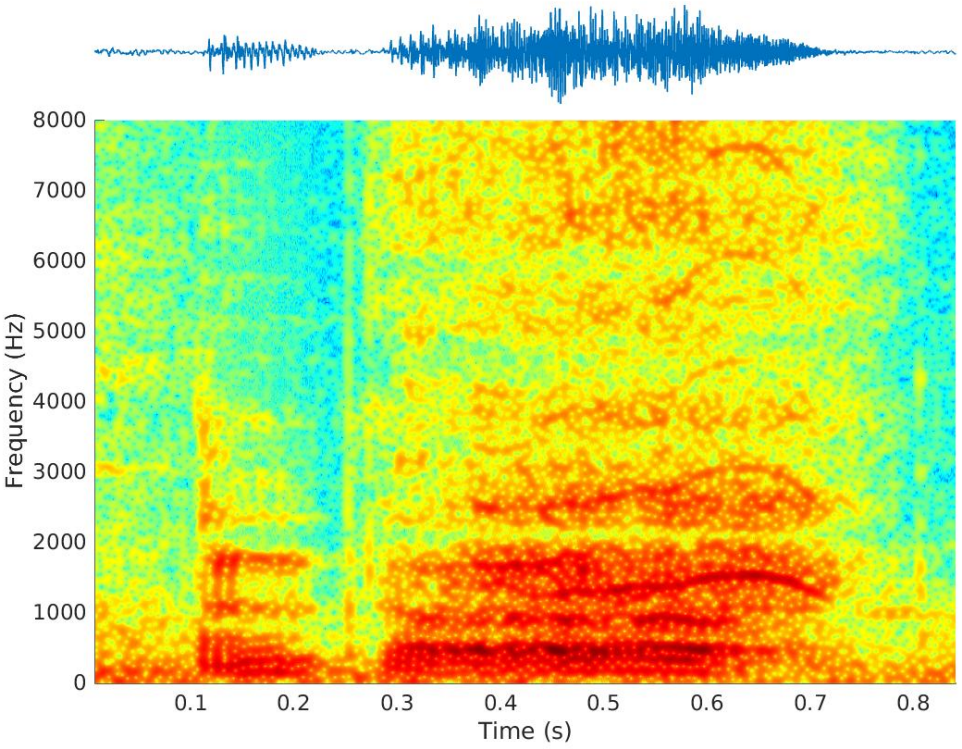

FIGURE 4.21: Spectrogram of the unprocessed vowel /u/

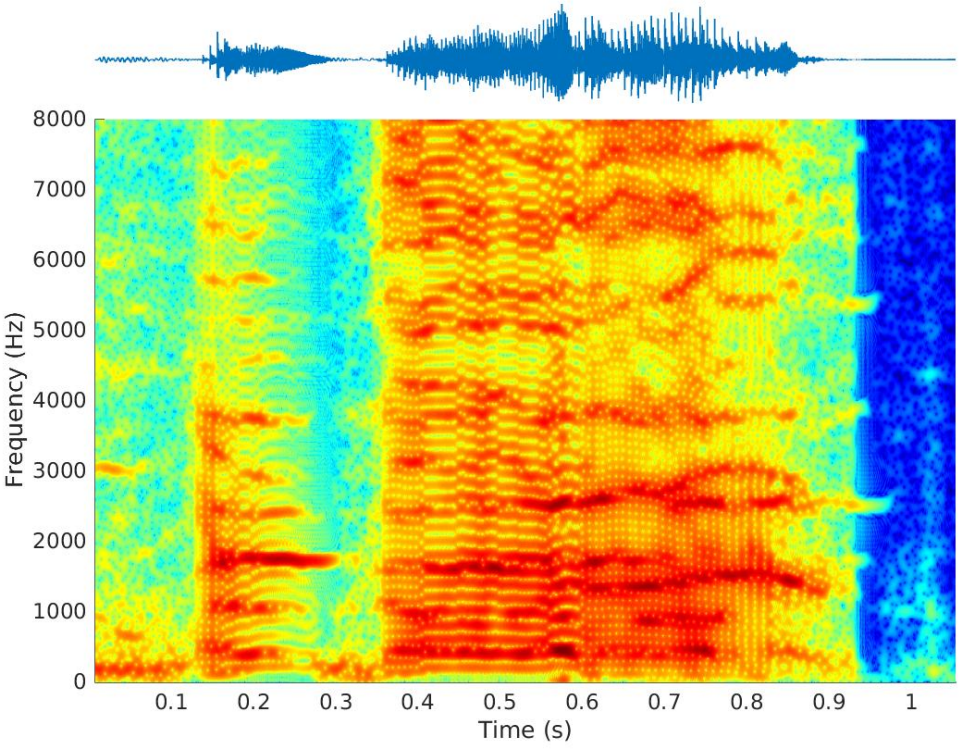

FIGURE 4.22: Spectrogram of the vowel /u/ processed with the proposed system

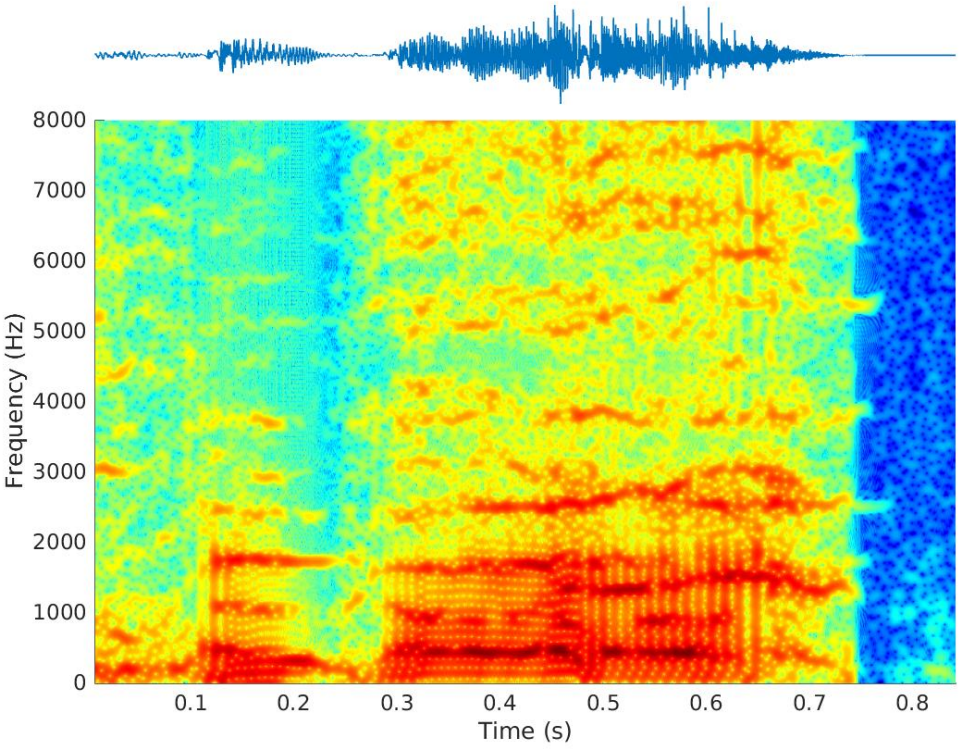

FIGURE 4.23: Spectrogram of the vowel /u/ processed with the reference system [83]

4.6 Chapter Summary

This chapter has provided the detailed results of the proposed system. The proposed system has been evaluated subjective listening test using MOS and preference score and objectively using HNR. The proposed system evaluation has been compared with the reference system [83]. The subjective listening test using MOS has shown that the proposed system, always provided good quality speech according to the MOS scaling (table 4.1). The preference test has shown that the proposed system preference all the time is more than the 50%. At the end the HNR has shown around 15 dB improvement over reference system and original speech samples. The results also has been visually inspected by spectrogram in this chapter.

Chapter 5

Conclusion

This thesis has presented and implemented the Esophageal Speech (ES) (a pathology speech used after total laryngectomy) enhancement method successfully. The thesis has implemented the signal processing algorithms to deal the main problems of ES regarding its source and vocal tract filter components. First the ES is decomposed into its source and vocal tract filter components using automatic inverse filtering method Iterative Adaptive Inverse Filtering (IAIF). The source and vocal tract filter are then transformed to normal speech components independently. The source signal is most effect component in ES has been transformed to normal speech source signal by natural glottal pulse by borrowing fundamental frequency F_0 curve and Harmonic to Noise Ratio (HNR) from the normal speech. The vocal tract filter problems which are spectral emphasis in higher frequencies, spectral peaks movement in higher frequencies and spectral peaks widths are corrected using the spectral de-emphasis, Frequency Warping Function (FWF) and spectral peaks bandwidth enlargement respectively.

The proposed system was evaluated subjectively using two listening tests, i) using Mean Opinion Score (MOS) and, ii) preference score in percentage. The objective assessment of the system was done using Harmonic to Noise Ratio (HNR). The proposed system was compared with reference system [83]. The improvement in MOS and the higher percentage of preference score have shown the capability of improving the intelligibility of ES in comparison to reference system and original speech samples. Furthermore the higher value of HNR objectively validated the proposed system capability, which resulted in higher quality ES.

5.1 Accomplished Objectives

Regarding the set objectives for this thesis;

- The high quality ES speech samples database has been recorded from the "Asociación Vizcaína de Laringectomizados y Mutilados de la Voz" of Bilbao. The database consists 28 two-syllable (CVCV) ES words, and sustained vowels (/a/, /e/, /i/, /o/, /u/). Every word and vowel has 18 utterances and database has total of 594 utterances of words and vowels. This large database has provided enough support for the proposed system research.
- The ES has been decomposed into its source and vocal tract filter components successfully, using the automatic inverse filtering Iterative Adaptive Inverse Filtering (IAIF).
- The transformed source signal which is near to normal speech signal, is obtained using the natural glottal pulse obtained from the natural speech. The original source spectrum and frame energy are used with normal speech fundamental frequency and Harmonic to Noise Ratio (HNR) for the modified source signal.
- The vocal tract filter has been transformed to normal speech vocal tract filter using spectral de-emphasis, Frequency Warping Function (FWF) and spectral peaks width enlargement.
- The processed ES has been evaluated objectively, and subjectively. Objectively, it is assessed using two listening tests, one MOS, and another preference score. Subjectively, proposed system is assessed using HNR.

5.2 Scientific Impact

The work presented in this thesis has following list of publications;

Journals

- Rizwan Ishaq, Begoña García Zapirain, " Optimal Subband Kalman Filter for Normal and Oesophageal Speech Enhancement", Bio-Medical Materials and Engineering, Vol. 24 (6), pp: 3569-3578, September 2014 [111]
- Rizwan Ishaq, Begoña García Zapirain, "Enhancement of Spanish Oesophageal Speech Vowels using Coherent Subband modulator Kalman Filtering", Technology and Health Care (Accepted).
- Rizwan Ishaq, Begoña García Zapirain, "Enhancement of Early Stage Spanish Esophageal Speech Using Modified Glottal Flow and Vocal Tract Using Natural

Pulse and Frequency Warping", Journal on Audio, Speech and Music Processing (in Review process)

Conferences

- Rizwan Ishaq, Begoña García Zapirain, "Adaptive Gain Equalizer for Improved of Esophageal Speech", IEEE international Symposium on Signal Processing and Information Technology (ISSPIT), pp: 153-157, Dec 2012 [177].
- Rizwan Ishaq, Muhammad Shahid, Benny Lovstrom, Begoña García Zapirain, Ingvar Claesson, " Modulation Frequency Domain Adaptive Gain Equalizer using Convex Optimzation", International Conference on Signal Processing and Communication Systems (ICSPCS), pp: 1-5, Dec 2012 [178].
- Rizwan Ishaq, Begoña García Zapirain, Muhammad Shahid, Benny Lovstrom, " Subband Modulator Kalman Filtering for Single Channel Speech Enhancement", IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), pp: 7442-7446, May 2013 [110].
- Rizwan Ishaq, Begoña García Zapirain, "Esophageal Speech Enhancement Using Modified Voicing Source", IEEE international Symposium on Signal Processing and Information Technology (ISSPIT), pp: 210-214, Dec. 2013 [179]
- Rizwan Ishaq, Dhananjaya, N Gowda, Paavo Alku, Begoña García Zapirain, "Vowel Enhancement in Early Stage Spanish Esophageal Speech Using Natural Glottal Flow Pulse and Vocal Tract Frequency Warping", 6th Workshop on Speech and Language Processing for Assisitive Technologies [180].

Others

- Rizwan Ishaq, Muhammad Shahid, Benny Sallberg, Benny Lovstrom, Nedelko Grbic, Ingvar Claesson, " Modulation Domain Adaptive Gain Equalizer for Speech Enhancement", ACTA Press, June 2011 [159].

5.3 Future Lines

Based on the proposed system, several future research lines can be established.

- The decomposition of voiced ES into source and vocal tract filter is good using IAIF, but still needs accurate and modified methods for decomposition by taking into account the vibration of esophagus which can be observed by Electro Glottograph (EGG) and high speed cameras.
- Based on this estimation of accurate source estimation, the vocal tract filter can be estimated perfectly, and then optimized algorithms can be designed
- In order to improve the social life of laryngectomee, the proposed system can be implemented to real world scenario, such as on digital signal processor devices, microcontroller or in mobile phone as a software which can be used to convert the ES into normal speech.
- The current implementing of proposed system work off-line i.e. database is already available. The proposed system in future can be test with on-line real time communication.

Appendix A

Publications

Vowel Enhancement in Early Stage Spanish Esophageal Speech Using Natural Glottal Flow Pulse and Vocal Tract Frequency Warping

Rizwan Ishaq[1]*, Dhananjaya Gowda*[2]*, Paavo Alku*[2]*, Begoña García Zapirain*[1]

[1]Deustotech-LIFE, University of Deusto, Bilbao, Spain
[2]Aalto University, Dept. of Signal Processing and Acoustics, Finland

rizwanishaq@deusto.es, dhananjaya.gowda@aalto.fi, paavo.alku@aalto.fi, mbgarciazapi@deusto.es

Abstract

This paper presents an enhancement system for early stage Spanish Esophageal Speech (ES) vowels. The system decomposes the input ES into neoglottal waveform and vocal tract filter components using Iterative Adaptive Inverse Filtering (IAIF). The neoglottal waveform is further decomposed into fundamental frequency F_0, Harmonic to Noise Ratio (HNR), and neoglottal source spectrum. The enhanced neoglottal source signal is constructed using a natural glottal flow pulse computed from real speech. The F_0 and HNR are replaced with natural speech F_0 and HNR. The vocal tract formant frequencies (spectral peaks) and bandwidths are smoothed, the formants are shifted downward using second order frequency warping polynomial and the bandwidth is increased to make it close to the natural speech. The system is evaluated using subjective listening tests on the Spanish ES vowels /a/, /e/, /i/, /o/, /u/. The Mean Opinion Score (MOS) shows significant improvement in the overall quality (naturalness and intelligibility) of the vowels.

Index Terms: speech enhancement, glottal flow, analysis synthesis vocal tract, spectral sharpening, warping

1. Introduction

The removal of the larynx after a Total Laryngectomy (TL), changes the speech production mechanism. The trachea which connects the larynx and lungs for air source is now connected to a stoma (hole on neck) for breathing. The vocal folds which resided in larynx are no more available. After TL, there is no voicing and air source for speech production. Therefore alternative voicing and air source are needed for speech restoration. Three methods are available for this purpose, i) Esophageal Speech (ES), ii) Tracheo-Esophageal Speech (TES), and iii) Electrolarynx (EL). ES and TES both use a common voicing source, the Phyarngo-Esophageal (PE) segment, but with a different air source, while EL uses external devices for voicing source with no air source. The ES is preferred over other methods, because it does not require surgery (TES) or external devices (EL). ES involves, however, a low pressure air source, and an irregular PE segment vibration which results in low quality and low intelligible speech. Compared to the production of normal speech according to the source-filter model [1], the voicing source in ES is severely altered and does not have any fundamental frequency or harmonic components. The vocal tract filter is also shortened in ES. The ES can be enhanced by transforming the source and filter components to those of normal speech using signal processing algorithms.

In previous studies ES is typically decomposed into its source and filter components using Linear Predication (LP) based analysis-synthesis techniques. Based on this assumption the authors in [2, 3] replaced the voicing source with the Liljencrants- Fant (LF) voicing source, and reported significant enhancements. Fundamental frequency smoothing and correction with the synthetic LF source model were used for quality enhancement also in [4]. ES enhancement based on formant synthesis has also shown significant improvement in intelligibility [5, 6]. In [7] the source and filter components were modified by replacing the source with the LF model and increasing the bandwidth of filter formants for better quality speech. Statistical conversion from ES to normal speech has also improved intelligibility, but requires more ES data [8]. Some other not so common approaches are based on Kalman filtering [9, 10, 11, 12], and modulation filtering enhancement [13, 14].

Almost all methods available in the literature assume that the fundamental frequency of ES can be estimated accurately. The voicing source signal is then modified with the synthetic LF model voicing source. The vocal tract formants are typically considered to be the same as in normal speech signals. In reality, however, the fundamental frequency of ES is highly irregular and the voicing source resembles whispered speech. Moreover, formants center frequencies are affected by the shortening of vocal tract length due to surgery. In order to deal with these deficiencies, this paper proposes an ES enhancement method based on the GlottHMM single pulse synthesis [15, 16, 17]. The system decomposes ES into neoglottal waveform and vocal tract filter components using Iterative Adaptive Inverse Filtering (IAIF) [18]. Natural glottal pulse extracted from real speech is used to construct the glottal waveform by borrowing F_0 curve and HNR from normal speech. The vocal tract filter is also modified by smoothing the spectral peaks and their bandwidths. The spectral peaks of the vocal tract filter are also moved to lower frequencies in order to compensate the rising of formant in ES. The formant bandwidths are also increased for better quality speech. The system is validated with Spanish Esophageal Vowels subjectively using the Mean Opinion Score (MOS). The paper in next section describes the system in detail. The subsequent sections contain results, discussion and finally conclusions.

2. System Description

The proposed system, shown in Figure 1, is divided into three main components, i) analysis, ii) transformation, and iii) synthesis. The analysis part decomposes the voiced speech frame into its source and filter components. The transformation provides the modified source and filter components. Finally the modified components are combined in the synthesis part to generate enhanced ES.

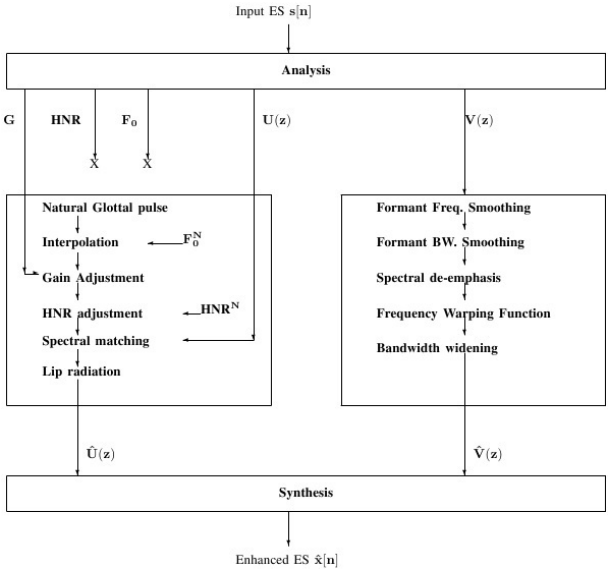

Figure 1: *Proposed enhancement system.*

2.1. GlottHMM based analysis

The goal of the analysis part of the system is to decompose the ES signal into a neoglottal source signal and a vocal tract spectrum. The input speech signal $s[n]$ is first passed through highpass filter $h_{hp}[n]$ with a cutoff frequency of 70 Hz.

$$s_h[n] = s[n] * h_{hp}[n] \tag{1}$$

where $s_h[n]$ and $*$ are the highpass filtered speech signal and a convolution operator, respectively. The highpass filtered signal $s_h[n]$ is then windowed using a rectangular window of size 45-ms, with 5-ms frame shift.

$$x[n] = s_h[n]w[n] \tag{2}$$

where $w[n]$ is the rectangular window. Firstly the log energy G of frame is extracted using,

$$G = log(\sum_{n=0}^{N-1} x^2[n]) \tag{3}$$

where N is the number of samples in the frame. Glottal Inverse Filtering (GIF) is then used to separate the frame into a neoglottal source signal and a vocal tract spectrum. The automatic inverse filtering, IAIF is used [18]. IAIF estimates vocal tract and lip radiation using all-pole modeling and then iteratively cancel these components. In simplified form, the neoglottal source signal:

$$U(z) = \frac{X(z)}{V(z)R(z)} \tag{4}$$

where $U(z)$, $X(z)$, $V(z)$ and $R(z)$ are the z-transforms of neoglottal source signal $u[n]$, speech signal $x[n]$, vocal tract impulse response $v[n]$, and lip radiation response $r[n]$ respectively. The estimated neoglottal source signal $u[n]$ is parametrized into fundamental frequency F_0, Harmonic to Noise Ratio (HNR) and neoglottal source spectrum $U(z)$. The autocorrelation of the neoglottal source signal $u[n]$ is used for F_0 estimation. The HNR is estimated using the upper and lower

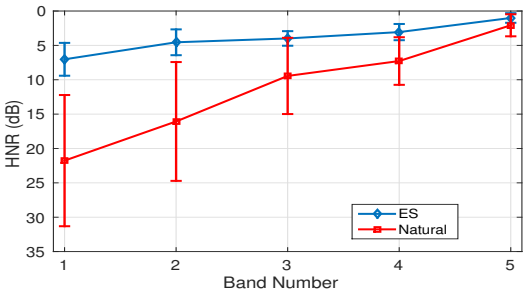

Figure 2: *HNR of ES and natural speech.*

smoothed spectral envelopes ratio to determine the voicing degree in the neoglottal voicing source signal $u[n]$ for five frequency bands [15]. In short the analysis part of the system provides for each frame the following, i) Frame energy G, ii) vocal tract spectrum $V(z)$ (LP order 30), iii) F_0, iv) HNR and v) neoglottal source spectrum $U(z)$ (LP order 10).

2.2. ES to normal speech transformation

The parameters obtained from the analysis are transformed into natural speech parameters. The neoglottal signal and vocal tract are modified independently.

2.2.1. Neoglottal source signal enhancement

The neoglottal source signal $u[n]$ is the most effected speech component in ES. Therefore the parameters of this signal are replaced with any arbitrary natural speech signal for a better glottal source signal. The natural glottal pulse which is extracted from normal speech is first interpolated using the cubic spline interpolation by replacing the frame original F_0 with natural speech F_0^N . The interpolated glottal pulse voicing source is then multiplied with the smooth gain G and the natural speech HNR is then used to add noise in the frequency domain for naturalness according to the following steps:

- Taking FFT of the neoglottal waveform,
- Adding random components (white Gaussian noise) to real and imaginary part of FFT according to HNR,
- Taking IFFT of noise added neoglottal waveform

$$U_{syn}(z) = 10^G G(z) + Q(z) \tag{5}$$

where $U_{syn}(z)$ is the synthetic glottal source, $G(z)$ is the natural glottal pulses source, and $Q(z)$ is HNR based noise component. Figure 2 shows the mean value of HNR for all voiced frames along with standard deviation. The figure indicates that HNR of ES is greatly different from that of normal speech. Therefore, it is justified to replace the HNR of ES with the HNR of normal speech in the vowel enhancement system. In order to adjust the spectrum of neoglottal waveform to the spectrum of the target waveform, the former is filtered with following IIR filter:

$$H_m(z) = \frac{U(z)}{U_{syn}(z)} \tag{6}$$

where $U(z)$ and $U_{syn}(z)$ are the LP spectra of the original and synthetic neoglottal waveform, respectively. The lip radiation is applied to the spectrally matched neoglottal waveform $\hat{u}[n]$:

$$\hat{u}[n] = \hat{u}[n] - \alpha\hat{u}[n-1], \quad 0.96 < \alpha < 1 \tag{7}$$

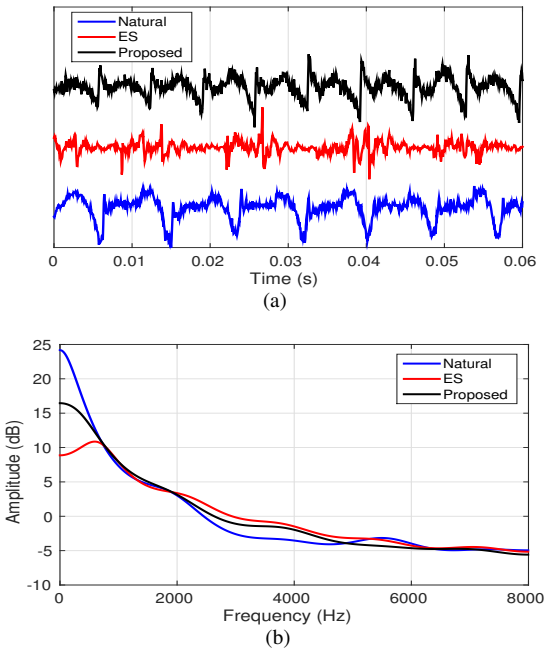

Figure 3: *Glottal excitations (computed from the vowel /a/) in the time domain (a) and in the frequency domain (b).*

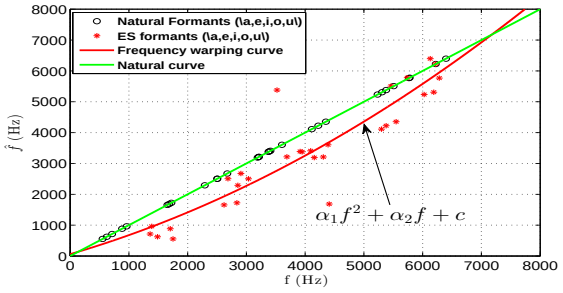

Figure 4: *Frequency Warping Function (FWF) curve.*

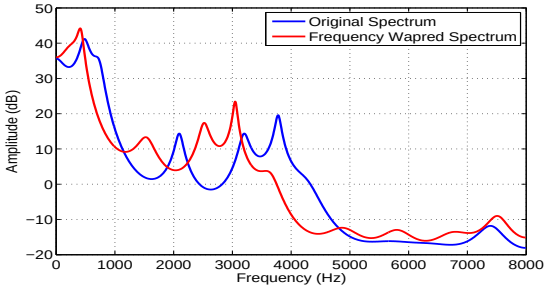

Figure 5: *Frequency warped spectra.*

where $\hat{u}[n](\hat{U}(z))$ and $\alpha(0.98)$ are the modified neoglottal waveform and lip radiation constant, respectively.

Figure 3(a) shows time-domain examples of glottal excitations of natural speech and ES together with a waveform computed with the proposed enhancement system. It can be seen that the proposed system is capable of producing a glottal excitation that is highly similar to that of natural speech. As shown in Figure 3(b), the spectral slope of the excitation waveform generated by the proposed method is also close to that of natural speech, especially at low frequencies, but the generated spectrum also retains the spectral slope of ES at higher frequencies.

2.2.2. *Vocal tract modification by nonlinear frequency warping*

The vocal tract spectrum of ES has the following characteristics, i) higher frequencies are emphasized more compared to lower frequencies, ii) spectral resonances (formants) are moved to higher frequencies, and iii) resonance bandwidths are reduced in comparison to normal speech vowels. To cope with the higher frequency emphasis, a de-emphasis filter is applied to the vocal tract spectrum. The resulting vocal tract transfer function is then expressed as:

$$H_{enh}(z) = \frac{1+\alpha z^{-1}}{1+\sum_{p=1}^{P} a_p z^{-p}}, \quad 0.95 < \alpha < 1 \qquad (8)$$

where P is the order of the all-pole vocal tract filter and α is the de-emphsis constant.

Because formants of ES are moved upward in frequency, a procedure is needed to adjust them to coincide more closely with the formant values of normal speech. For such a procedure, we used a second order Frequency Warping Function (FWF) $\zeta(f)$ defined as:

$$\zeta(f) = \alpha_1 f^2 + \alpha_2 f + c \qquad (9)$$

where $\alpha_1 = 6.079 \times 10^{-5}$, $\alpha_2 = 0.5553$, and $c = 60.280$.

$$\hat{f} = \beta\zeta(f), \quad \beta = 1, f = 0 \rightarrow \frac{f_s}{2} \qquad (10)$$

where $\hat{f}$ and f, are warped and original frequencies, and β is a constant. Figure 4 demonstrates FWF using first four formants of vowels (/a/, /e/, /i/, /o/, /u/) extracted from normal speech (x-axis) and ES (y-axis). The obtained frequency warping, applicable for a general formant mapping between normal speech and ES, is shown in Figure 5. In order to expand the formant bandwidths, exponential windowing is used for the vocal tract filter coefficients as follows [19]:

$$H_s(z) = \frac{1+\sum_{p=1}^{P} \gamma^p a_p z^{-p}}{1+\sum_{p=1}^{P} \eta^p a_p z^{-p}}, \quad 0.90 < \gamma, \eta < 1 \qquad (11)$$

where γ and η are constants controlling the spectral bandwidth.

If $\gamma > \eta$ bandwidth of formants increase, otherwise it decreases (i.e. formants are sharpened). For the purpose of the present study, $\eta(0.97)$ is always smaller than $\gamma(0.99)$ in order to increase formant bandwidths.

2.3. Synthesis of enhanced speech

The synthesis part involves convolving the modified neoglottal waveform and the impulse response of the vocal tract filter yielding the enhanced version of ES $\hat{x}[n]$;

$$\hat{x}[n] = \hat{v}[n] * \hat{u}[n] \qquad (12)$$

where $\hat{u}[n]$ and $\hat{v}[n]$ are the modified neoglottal waveform and vocal tract impulse response, respectively.

3. System Evaluation

The system was evaluated with ES vowels of Spanish (/a/, /e/, /i/, /o/, /u/) recorded in speech rehabilitation center. The data

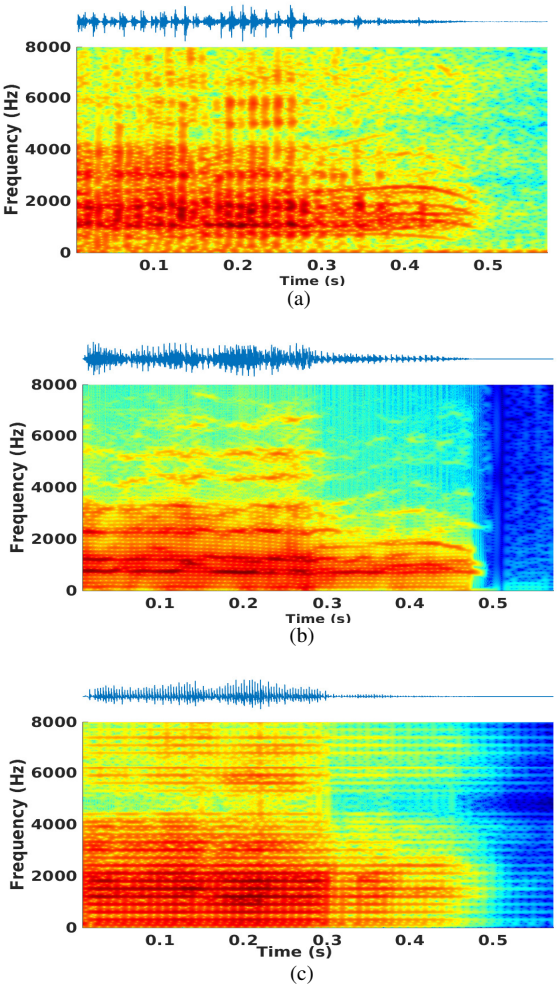

Figure 6: *Spectrograms of the vowel /a/ for different processing types: unprocessed (a), processed with the proposed system (b), processed with the reference system (c) [7]*

was collected from five early stage male ES talkers by asking them to utter each vowel four times. Due to lack of female patients in the rehabilitation center, only male speakers were involved in the study. The speech sounds were sampled with 44.1 kHz from which the data was down-sampled to 16 kHz for computational efficiency.

The system performance is visually demonstrated with spectrograms in Figure 6. In this figure, and also later in Figures 7 and 8, the proposed system is compared with a reference system based on using the LF source and formant modification with a bandwidth extension system [7]. It can be seen from Figure 6 that the spectrogram computed from the enhanced vowels by the proposed system shows a clearer formant and harmonics structure in comparison to ES and the reference system.

3.1. Subjective listening evaluation

Two subjective listening tests were conducted. The first one was a quality evaluation based on the Mean Opinion Score (MOS) which is a widely used perceptual quality test of speech based on a scale from 1 (worst) to 5 (best). In this test, the listeners heard original ES vowels and the corresponding enhanced ones,

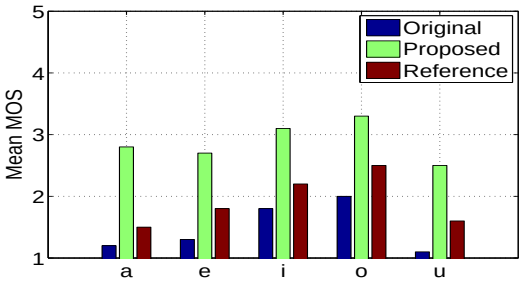

Figure 7: *Results of the MOS test for all the vowels.*

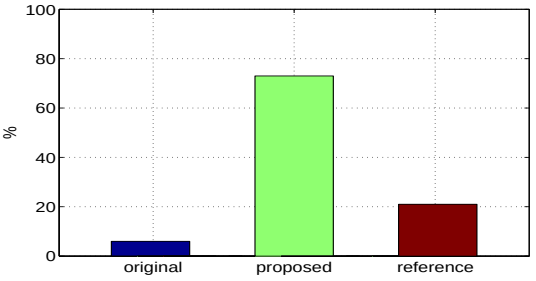

Figure 8: *Results of the preference test.*

processed by both the proposed and the reference method, in a random order and they were asked to grade the quality of the sounds on the MOS scale. The second listening test was a preference test where the listeners heard vowels corresponding to the same three processing types and they were asked to select which one they prefer to listen. A total of 10 listeners participated in the listening tests.

Figure 7 shows the results of the MOS test. The data indicates that the proposed system has a mean MOS higher than 2.5 for all the vowels, which can be considered a good quality score for ES samples. Figure 8 shows the data of the preference tests by combining all the vowels. Also these data indicate that the proposed method has succeeded in enhancing the quality of the ES vowels.

4. Conclusion

An enhancement system for ES vowels was proposed based on using a natural glottal pulse combined with second order polynomial Frequency Warping Function. A preliminary evaluation of the system was carried out on early stage Spanish ES vowels by comparing the system performance with a known reference method. Results obtained with a MOS evaluation show clear improvements in speech quality both in comparison to the original ES vowels and to sounds enhanced with the reference method. The good performance was corroborated with a preference test indicating that in the vast majority of the cases, listeners preferred to listen to the sounds enhanced by the proposed method. Future work is needed to study the system together with advanced stage ES speakers.

5. Acknowledgements

Special thanks to all my colleagues at Aalto University for their valuable support and time.

6. References

[1] G. Fant, "Acoustic theory of speech production." Mouton, The Hauge, 1960.

[2] Q. Yingyong, W. Bernd, and B. Ning, "Enhancement of female esophageal and tracheoesophageal speech," *Acoustical Society of America*, vol. 98(5, Pt1), pp. 2461–2465, 1995.

[3] Y. Qi, "Replacing tracheoesophageal voicing source using lpc synthesis," *Acoustical Society of America*, vol. 5, pp. 1228–1235, 1990.

[4] R. Sirichokswad, P. Boonpramuk, N. Kasemkosin, P. Chanyagorn, W. Charoensuk, and H. H. Szu, "Improvement of esophageal speech using lpc and lf model," *Internation Conf. on Biomedical and Pharamaceutical Engineering 2006*, pp. 405–408, 2006.

[5] M. Kenji, H. Noriyo, K. Noriko, and H. Hajime, "Enhancement of esophageal speech using formant synthesis," *Acoustic. Sci. and Tech.*, pp. 69–76, 2002.

[6] M. Kenji and H. Noriyo, "Enhancement of esophageal speech using formant synthesis," *Acoustics, Speech and Signal Processing, International conf.*, pp. 81–85, 1999.

[7] R. H. Ali and S. B. Jebara, "Esophageal speech enhancement using excitation source synthesis and formant structure modification," *SITIS*, pp. 615–624, 2006.

[8] K. Doi, H.and Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Statistical approach to enhancing esophageal speech based on gaussian mixture models," *Acoustics Speech and Signal Processing(ICASSP), 2010 IEEE International Conference*, pp. 4250–4253, 2010.

[9] O. Ibon, B. Garcia, and Z. M. Amaia, "New approach for oesophageal speech enhancement," *10th International conference, ISSPA*, vol. 5, pp. 225–228, 2010.

[10] B. Garcia and A. Mendez, "Oesophageal speech enhancement using poles stablization and kalman filtering," *ICASSP*, pp. 1597–1600, 2008.

[11] B. Garcia, I. Ruiz, A. Mendez, and M. Mendezona, "Oesophageal voice acoustic parameterization by means of optimum shimmer calculation," *WSEAS Trasactions on Systems*, pp. 489–499, 2008.

[12] R. Ishaq and B. G. Zapirain, "Optimal subband kalman filter for normal and oesophageal speech enhancement," *Bio-Medical Materials and Engineering*, vol. 24, pp. 3569–3578, 2014.

[13] R. Ishaq, B. G. Zapirain, M. Shahid, and B. Lovstrom, "Subband modulator kalman filtering for signla channel speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[14] R. Ishaq and B. G. Zapirain, "Adaptive gain equalizer for improvement of esophageal speech," in *IEEE International Symposium on Signal Processing and Information Technology*, 2012.

[15] A. Suni, T. Raitio, , M. Vainio, and P. Alku, "The glottalHMM entery for blizzard challenge 2011: Utilizing source unit selection in hmm-based speech synthesis for improved excitation generation," in *in Blizzard Challenge 2011, Workshop, Florence, Italy*, 2011.

[16] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, pp. 153–165, 2011.

[17] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "Utilizing glottal source pulse library for generating improved excitation signal for hmm-based speech synthesis," in *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2011.

[18] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," in *Speech communication*, vol. 11, no. 2, 1992, pp. 109–118.

[19] J. H. Chen and A. Gersho, "Adaptive postfiltering for quality enhancement of coded speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, pp. 59–71, 1995.

Bio-Medical Materials and Engineering 24 (2014) 3569–3578
DOI 10.3233/BME-141183
IOS Press

Optimal subband Kalman filter for normal and oesophageal speech enhancement

Rizwan Ishaq* and Begoña García Zapirain
Deustotech-LIFE, University of Deusto, Bilbao, Spain

Abstract. This paper presents the single channel speech enhancement system using subband Kalman filtering by estimating optimal Autoregressive (AR) coefficients and variance for speech and noise, using Weighted Linear Prediction (WLP) and Noise Weighting Function (NWF). The system is applied for normal and Oesophageal speech signals. The method is evaluated by Perceptual Evaluation of Speech Quality (PESQ) score and Signal to Noise Ratio (SNR) improvement for normal speech and Harmonic to Noise Ratio (HNR) for Oesophageal Speech (OES). Compared with previous systems, the normal speech indicates 30% increase in PESQ score, 4 dB SNR improvement and OES shows 3 dB HNR improvement.

Keywords: Kalman filter, autoregressive, speech enhancement, weighted linear prediction

1. Introduction

The Kalman filter is considered optimal among other signal enhancement methods, such as Wiener filtering, spectral subtraction, wavelet denoising, etc. [1–3]. Inheritance of speech production model and non-stationary signal processing are the advantages of Kalman filtering over other speech enhancement methods [4]. The Kalman filter is first introduced for speech enhancement by Paliwal, providing clean speech Autoregressive (AR) coefficients, and noise variances using conventional Linear Prediction (LP) [5]. The further modification to [5] is done by estimating AR coefficients recursively through Expectation Maximization (EM) algorithms [6] and modeling colored noise as AR process [7–9]. The Kalman filter is used in frequency subbands for efficient and low complex processing with fewer number of AR coefficients using conventional LP [10–12].

The Oesophageal Speech (OES) is a special type of alaryngeal speech used after the treatment of laryngeal cancer for rehabilitation of voice. Several techniques are available for speech restoration, most important being Oesophageal Speech (OES), Tracheo-Esophageal Speech (TES) and Electrolarynx (EL) speech. The OES is most used method because it requires no external device (EL) and surgery (TES). Despite its advantages, OES has low fundamental frequency and intelligibility, due to irregular vibration of esophagus and noise. The Kalman filter is also utilized to enhance the quality of OES such as, fullband Kalman filtering [13–15] and subband Kalman filtering [12]. Kalman filtering has shown significant enhancement over other methods such as source-filter decomposition [16] for enhancing source and filter [17–20], use of LF voicing source [21,22], and statistical methods [23].

*Corresponding author: Rizwan Ishaq, Deustotech-LIFE, University of Deusto, Bilbao, Spain. Tel.: +34 94 413 90 03; Fax: +34 94 413 90 03; E-mail: rizwanishaq@deusto.es.

Speech enhancement using Kalman filter needs optimal speech, noise AR coefficients and variances. Previous methods [4–7,12,13] use conventional LP for this purpose, but the conventional LP has the sensitivity problem with additive background noise [24]. The estimation of noise AR coefficients also needs non-speech activity detector. To overcome these problems, this paper presents the optimal subband Kalman filter by providing optimum AR coefficients for speech and noise signals usingWeighted Linear Prediction (WLP) [24] and Noise Weighting Function (NWF) [25]. The proposed system is evaluated, for normal speech by Perceptual Evaluation of Speech Quality (PESQ) score and Signal to Noise Ratio (SNR) improvement, and for OES using Harmonic to Noise Ratio (HNR). The paper outline is as follow; Section 2 provides the detailed description of the system, Section 3 provides the optimal parameter estimation, followed by simulation results in Section 4 and Conclusion in Section 5.

2. System design

The proposed system (KF-P) components are shown in Figure 1, and subsequent sections provide the detail of every component.

2.1. Analysis filter bank

The analysis filterbank is used to decompose the input speech signal x(n) into different subbands. The x(n) passes through the filterbank of different Linear Time Invariant (LTI) bandpass filters, each having impulse response of $h_k(n)$, mathematically the frequency subband signal is [26]:

$$x_k(n) = x(n) * h_{(k)(n)} \tag{1}$$

where * is the convolution operator.

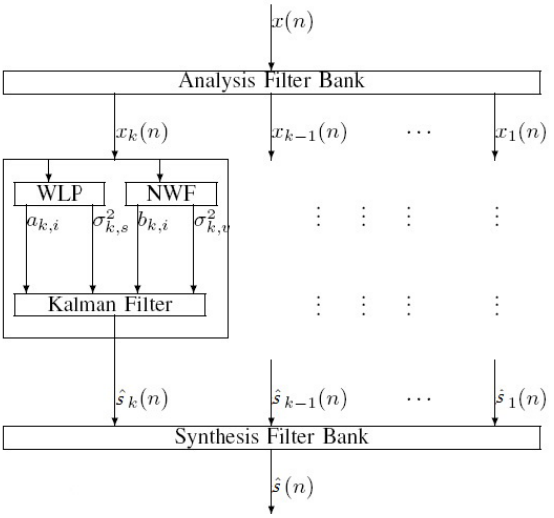

Fig. 1. Block diagram of proposed system.

2.2. Optimal subband Kalman filter

The subband signal $x_k(n)$ consists of clean speech $s_k(n)$ and noise $v_k(n)$ as follows:

$$x_k(n) = s_k(n) + v_k(n) \tag{2}$$

Both $s_k(n)$ and $v_k(n)$ can be modelled as AR processes of order p and q:

$$s_k(n) = \sum_{i=1}^{p} a_{k,i}(n) s_k(n-i) + \omega_k(n) \tag{3}$$

$$v_k(n) = \sum_{i=1}^{q} b_{k,i}(n) v_{(k)(n-i)} + \Gamma_k(n) \tag{4}$$

where $\omega_k(n)$ and $\Gamma_k(n)$ are uncorrelated additive white Gaussian noises with zero mean and variances $\sigma_{k,\omega}^2$ and $\sigma_{k,v}^2$ respectively. Given $s_k(n) = [s_k(n), \dots, s_k(n-p+1)]$, and $v_k(n) = [v_{(k)}, \dots, v_k(n-q+1)]$, above equations can be written in state-space domain for Kalman filtering:

$$s_k(n) = F_{k,s} s_k(n-1) + G_{k,s} \omega_k(n) \tag{5}$$

$$x_{k,s}(n) = H_{k,s}^T s_k(n) \tag{6}$$

$$v_k(n) = F_{k,v} v_k(n-1) + G_{k,v} \Gamma_k(n) \tag{7}$$

$$x_{k,v}(n) = H_{k,v}^T v_k(n) \tag{8}$$

where state transition matrices $F_{k,s}$ and $F\_(k,v)$ are :

$$F_{k,s} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ -a_{k,p} & -a_{k,p-1} & -a_{k,p-2} & \dots & -a_{k,1} \end{bmatrix}_{p,p} \tag{9}$$

$$F_{k,v} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ -b_{k,q} & -ba_{k,q-1} & -b_{k,q-2} & \dots & -b_{k,1} \end{bmatrix}_{q,q} \tag{10}$$

and $G_{k,s}, G_{k,v}, H_{k,s}$ and $H\_(k,v)$ are:

$$G_{k,s} = H_{k,s}^T = [0, 0, \dots, 1]_{px1}, G_{k,v} = H_{k,v}^T = [0, 0, \dots, 1]_{qx1}$$

Combining the above equations we have:

$$\hat{x}_k(n) = F_k\, \hat{x}_k(n-1) + G_k\, \widehat{\omega}_k(n) \tag{11}$$

$$x_k(n) = H_k^T\, \hat{x}_{(k)(n)} \tag{12}$$

where $\hat{x}_k(n), F_k, G_k, H_k,$ and $\widehat{\omega}_k(n)$ are given as:

$$\hat{x}_k(n) = [s_k(n), \dots, s_k(n-p+1), v_k, \dots, v_k(n-q+1)]$$

$$F_k = \begin{bmatrix} F_{k,s} & 0 \\ 0 & F_{k,v} \end{bmatrix}, G_k = \begin{bmatrix} G_{k,s} & 0 \\ 0 & G_{k,v} \end{bmatrix}$$

$$H_k^T = \left[H_{k,s}, H_{k,v}\right], \qquad \widehat{\omega}_k^T(n) = [\sigma_{k,s}^2, \sigma_{k,s}^2]$$

The Kalman filter provides optimal estimate $\hat{s}_k(n)$ of $x_k(n)$, with Kalman gain $K_k(n)$, prediction error covariance $P_k(n|n-1)$ and filtering error covariance $P_k(n)$ [5,12] :

$$\hat{\hat{x}}_k(n) = F_{k\,\hat{\hat{x}}_k(n-1)} + K_k(n)\left[x_k(n) - H_k^T F_k \hat{\hat{x}}_k(n-)\right] \tag{13}$$

$$K_k(n) = P_k(n|n-1)H_k^T\left[H_k P_k(n|n-1)H_k^T\right]^{-1} \tag{14}$$

$$P_k(n|n-1) = F_k P_k(n-1)F_k^T + G_k Q_k(n) G_k^T \tag{15}$$

$$P_k(n) = \left[\, I - K_k(n)H_k^T\right]P_k(n|n-1) \tag{16}$$

Where $\hat{\hat{x}}_k(n)$ is the estimated state vector. The covariance matrix of $\widehat{\omega}_k(n)$ is:

$$Q_k(n) = E\left\{\, \widehat{\omega}_k(n)\, \widehat{\omega}_k^T(n)\right\} = diag\left(\sigma_{k,s}^2, \sigma_{k,v}^2\right)$$

The desired $\hat{s}_k(n)$ is given as:

$$\hat{s}_k(n) = H_{k,s}^T \hat{\hat{x}}_k(n) \tag{17}$$

2.3. Synthesis filter bank

The filterbank summation method is used to reconstruct the enhanced fullband signal $\hat{s}(n)$, using the modified frequency subbands $\hat{s}_k(n)$ [26]:

$$\hat{s}(n) = \sum_{k=1}^{K} \hat{s}_k(n) \tag{18}$$

3. Optimal parameter estimation

The optimal estimation of speech, noise AR coefficients $a_{k,i}, b_{k,i}$ and their respective variances $\sigma_{k,s}^2, \sigma_{k,v}^2$ are necessary for the optimal results of Kalman filtering. This section provides the optimal AR coefficients and variance for speech using WLP [27], and colored noise optimal parameters by computing noise signal through NWF [25].

3.1. Weighted Linear Prediction (WLP)

The conventional Linear Prediction (LP) sensitivity to additive background noise produces poor AR coefficients [27]. The Conventional LP AR coefficients are estimated using Minimum Mean Square Error (MMSE) criterion. The prediction error $\varepsilon_k^2(n)$ partial derivative with respect to AR coefficients $a_{k,i}$ is set to zero [16,27]:

$$\frac{\partial}{\partial a_{k,i}} \varepsilon_k^2(n) = \frac{\partial}{\partial a_{k,i}} \left(\sum_n (x_k(n) - \sum_{l=1}^{p} a_{k,l} x_k(n-l))^2 \right) = 0 \tag{19}$$

Considering $r_{k,x}(l,i) = \sum_n x_k(n-l)x_k(n-i)$, the solution to Eq. (19) is:

$$\sum_{i=1}^{p} a_{k,i} r_{k,x}(l,i) = r_{k,x}(l,0), \;\; l = 1,2,\dots,p \tag{20}$$

To overcome the sensitivity of LP to additive background noise, WLP introduced weighting function $\zeta_k(n)$ using the Short Time Energy (STE) of size M, which provides better estimation of AR coefficients by focusing on high SNR region [24]:

$$\zeta_k(n) = \sum_{i=1}^{M} x_k^2(n-i) \tag{21}$$

Using $\zeta_k(n)$, the AR coefficients can be obtained, considering $r_{k,x}(l,i) = \sum_n \zeta_k(n) x_k(n-l)x_k(n-i)$:

$$\begin{bmatrix} a_{k,1} \\ a_{k,2} \\ \vdots \\ a_{k,p} \end{bmatrix} = \begin{bmatrix} r_{k,x}(1,1) & r_{k,x}(1,2) & \dots & r_{k,x}(1,p) \\ r_{k,x}(2,1) & r_{k,x}(2,2) & \dots & r_{k,x}(2,p) \\ \vdots & \vdots & \vdots & \vdots \\ r_{k,x}(p,1) & r_{k,x}(p,2) & \dots & r_{k,x}(p,p) \end{bmatrix}^{-1} \begin{bmatrix} r_{k,x}(1,0) \\ r_{k,x}(2,0) \\ \vdots \\ r_{k,x}(p,0) \end{bmatrix} \tag{22}$$

The speech variance $\sigma_{k,s}^2$ is [28]:

$$\sigma_{k,s}^2 = r_{k,x}(0,0) - \sum_{i=1}^{p} a_{k,i} r_{k,x}(0,i) \tag{23}$$

3.2. Noise Weighting Function (NWF)

The noise signal $v_k(n)$ is necessary for the sub-optimum AR coefficients $b_{k,i}$ and noise variance $\sigma_{k,v}$. The NWF estimates $v_k(n)$ using the ratio of noise to speech $\kappa_k(n)$ [25]:

$$\kappa_k(n) = \min\left\{\left(\frac{\gamma_k(n)}{\eta_k(n)+\epsilon}\right)^{\rho_k}, L_k\right\} \tag{24}$$

where ρ_k and L_k are gain rise exponent and limiting factor respectively. The $\eta_k(n)$ and $\gamma_k(n)$ are:

$$\eta_k(n) = \alpha_k \eta_k(n-1) + (1-\alpha_k)|x_k(n)| \tag{25}$$

$$\gamma_k(n) = \begin{cases} \eta_k(n) & if\ \eta_k(n) \le \gamma_{(k)}(n-1) \\ (1+\beta_k)\big(\gamma_k(n-1)\big) & otherwise \end{cases} \tag{26}$$

where $\alpha_k = \frac{1}{f_{k,s}T_{k,a}}$ and $\beta_k \frac{1}{f_{(k,s)}T_{k,b}}$ are forgetting factor constant and positive constant respectively. The $f_{k,s}$ is subband sampling frequency, and $T_{k,b}$ and $T_{k,a}$ are time constants controlling the noise level. The noise signal $v_k(n)$ is obtained by multiplying $\kappa_k(n)$ to $x_k(n)$:

$$v_k(n) = \kappa_k(n).x_k(n) \tag{27}$$

The AR coefficients $b_{k,i}$ can be calculated by solving following equation:

$$\sum_{i=1}^{q} b_{k,i} r_{k,v}(l,i) = r_{k,v}(l,0), \quad l = 1,2,\dots,q \tag{28}$$

Where $r_{k,v}(l,i) = \sum_n v_k(n-l)v_k(n-i)$. The noise variance $\sigma_{k,v}^2$ is [28]:

$$\sigma_{k,v}^2 = r_{k,v}(0,0) - \sum_{i=1}^{q} b_{(k,i)r_{k,v}(0,i)} \tag{29}$$

4. Evaluation results

4.1. Speech material and system setting

The performance of proposed system (KF-P) for normal speech, is tested by male (5 speakers) and female (5 speakers) speech signals of sampling frequency 16000 Hz [29], which are corrupted by Factory Noise (FN) and Engine Noise (EN) at different Signal to Noise Ratio (SNR) levels (-10, -5, 0, 5 ,10 dB). For OES, Spanish OES vowels \a\, \e\, \i\, \o\, and \u\are used, which are recorded from pa-

thology speech rehabilitation association (6 male speakers uttered each vowel 3 times and center does not have any female speaker).

The system uses 16 subbands filterbank [29]. Each subband uses segment and overlap size of 30 ms and 15 ms respectively. The prediction orders p and q for speech and noise are 12 and 6 respectively.

The proposed system (KF-P) for normal speech, is compared with following available subband based

Kalman filtering algorithms:

– (KF-1){subband Kalman filtering with optimal parameter estimated recursively using conventional
– Linear prediction [28]};
– (KAF-2){Subband Kalman filtering, where parameters estimated in modulation domain with conventional linear prediction [12]};
– (KF-3){ subband Kalman filtering where speech and noise parameters estimated using LMS algorithm [10]}.

The proposed algorithm (KF-P) for OES, is compared with following systems:

– (KF-1){fullband Kalman filter originally used for OES [13]};
– (KF-2){Subband Kalman filter in modulation domain using conventional linear prediction [12]};
– (KF-O){modification to [13] using poles stabilization [15]}.

4.2. PESQ

Table 1 and 2 show the PESQ scores for female and male speech, corrupted by engine noise and factory noise. The proposed system (KF-P) outperforms all other available subband Kalman filtering methods, particularly at low SNR.

4.3. Signal to Noise Ratio (SNR) improvement

The Table 3 and 4 has shown SNR improvement for male and female speech signals corrupted by factory and engine noises. The proposed system (KF-P) has shown around 3-4 dB improvement over the other methods.

Table 1

PESQ Scores for Engine Noise Corrupted Speech

SNR(dB)	**-10**	**-5**	**0**	**5**	**10**
Original	0.52	0.90	1.73	1.84	2.48
KF-P	1.34	1.48	2.53	3.18	3.91
KF-1	0.93	1.02	2.48	2.78	3.19
KAF-2	0.78	0.99	1.38	2.72	3.62
KF-3	0.28	0.69	1.59	2.81	3.71
Male Speaker					
Original	0.65	0.80	1.53	1.94	2.78
KF-P	1.17	1.67	2.83	3.28	3.47
KF-1	0.91	1.22	2.18	2.98	3.09
KAF-2	0.96	1.19	2.07	2.52	3.10
KF-3	1.03	1.07	2.09	2.91	3.26
Female Speaker					

Table 2

PESQ Scores for Factory Noise Corrupted Speech

SNR(dB)	**-10**	**-5**	**0**	**5**	**10**
Original	0.32	0.89	1.63	2.43	2.97
KF-P	1.20	1.68	2.03	2.78	3.70
KF-1	0.34	0.91	1.98	2.12	2.99
KAF-2	0.98	1.10	1.78	2.52	3.12
KF-3	0.78	0.99	1.89	2.11	3.31
Male Speaker					
Original	0.42	0.79	1.53	2.63	2.45
KF-P	1.40	1.88	2.53	3.21	3.50
KF-1	0.67	1.01	2.11	2.12	2.79
KAF-2	0.98	1.27	2.18	2.75	3.12
KF-3	1.18	1.69	2.39	2.94	3.01
Female Speaker					

Table 3

Signal to Noise Ratio (SNR) improvement for Engine Noise Corrupted Speech

SNR(dB)	**-10**	**-5**	**0**	**5**	**10**
KF-P	2.231	3.31	2.84	5.12	8.97
KF-1	0.33	1.98	2.03	3.98	7.3
KAF-2	1.94	2.21	2.38	4.12	7.09
KF-3	0.98	1.89	2.78	3.52	7.12
Male Speaker					
KF-P	3.342	4.42	3.95	6.23	9.89
KF-1	1.44	2.99	3.14	4.99	6.41
KAF-2	2.83	3.32	3.49	4.23	6.10
KF-3	1.09	2.90	3.89	4.63	5.23
Female Speaker					

Table 4

Signal to Noise Ratio (SNR) improvement for Factory Noise Corrupted Speech

SNR(dB)	**-10**	**-5**	**0**	**5**	**10**
KF-P	2.321	4.61	5.12	7.02	9.17
KF-1	1.33	2.98	3.03	4.98	5.3
KAF-2	1.54	2.51	3.38	4.52	6.09
KF-3	1.98	2.89	4.78	5.52	7.12
Male Speaker					
KF-P	2.34	4.12	4.95	8.23	9.90
KF-1	1.54	2.89	3.45	5.69	7.21
KAF-2	1.93	3.12	3.19	5.13	6.10
KF-3	1.99	3.90	4.19	6.13	7.93
Female Speaker					

4.4. Harmonic to Noise Ratio (HNR)

The HNR parameter is used, extensively by OES research community for quality measurement [30]. The HNR parameter is calculated using the freely available speech analysis software VoiceSauce [31], according to following settings: segment length and overlap are 30 milliseconds and 15 milliseconds respectively, fundamental frequency estimated using STRAIGHT [32] method in the range of 50 to

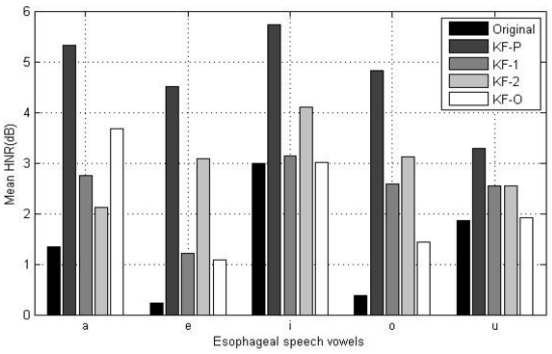

Fig. 2. Mean harmonic to Noise Ratio (HNR) for Spanish vowels.

120 Hz (OES fundamental frequency falls in this range), and prediction order is set to 12. The Figure 2 has shown the mean HNR improvement of 3dB over previously available methods, particularly for vowels \a\, \i\.

5. Conclusion

The system successfully implemented the subband Kalman filtering, by providing optimized parameters for speech and noise. The optimal AR coefficients for speech and its variance are estimated utilizing Weighted Linear Predication (WLP). On the other hand, noise AR coefficients are estimated by calculating Noise Weighting Function (NWF) for subband signals. The method has outperformed all the available subband Kalman filtering both for normal and Oesophageal Speech (OES) signals objectively. The normal speech has shown its superiority through improved PESQ scores and SNR improvement, while for OES signals, improvement has been shown by improved HNR.

Acknowledgement

This research was granted by Deiker of Deusto University and Department of Eduction and Research of Basque government.

References

[1] M.H. Hayes, Statistical Digital Signal Processing and Modelling, John Wiley & Sons, Inc., New York, USA, 1st edi., 1996.
[2] Z. Ji and F. Wang, Application of the dual-tree complex wavelet transform in biomedical signal denoising, Bio-Medical Materials and Engineering **24** (2014), 109–115.
[3] B.P. Hu, Y. Li, D.Y. Qiao and T. He, Non-contact physiological signal detection using continuous wave doppler radar, Bio-Medical Materials and Engineering **24** (2014), 993–1000.
[4] S. So and K.K. Paliwal, Suppressing the influence of additive noise on the kalman gain for low residual noisespeech enhancement, Elsever, Speech Communication **53** (2011), 355–378.
[5] K.K. Paliwal and A. Basu, A speech enhancement method based on Kalman filtering, IEEE Int. Conf. Acoust., Speech, Signal Processing **12** (1987), 177–180.

[6] D. Weixiu and P. Driessen, Speech enhancement based on Kalman filtering and EM algorithm, IEEE Pacific. Rim. Conf. on Communication, Computers and Signal Processing **1** (1991), 142–145.
[7] D.J.G.B. Koo and S.D. Gray, Filtering of colored noise for speech enhancement and coding, IEEE Trans. on Signal Processing **8** (1991), 1732–1742.
[8] D.C. Popescu and I. Zeljkovic, Kalman filtering of colored noise for speech enhancement, Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing **2** (1998), 997–1000.
[9] H. Puder, Kalman-filter in subbands for noise reduction with enhanced pitch-adaptive speech model estimation, European Transactions on Telecommunications **13** (2002), 139–148
[10] W.-R. Wu and P.-C. Chen, Subband Kalman filtering for speech enhancement, circuits and systems II, IEEE Transactions on Analog and Digital Signal Processing **45** (1998), 1072–1083.
[11] S. So and K.K. Paliwal, Modulation-domain kalman filtering for single-channel speech enhancement, Speech Commun. **53** (2011), 818–829.
[12] R. Ishaq, B.G. Zapirain, M. Shahid and B. Lovstrom, Subband modulator kalman filtering for single channel speech enhancement, IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, 7442–7446.
[13] B. Garcia and A. Mendez, Oesophageal speech enhancement using poles stablization and Kalman filtering, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2008, 1597–1600.
[14] B. Garcia, J. Vicente, A. Alonso and E. Loyo, Esophageal voices: Glottal flow restoration, Acoustics, Speech and Signal Processing **4** (2005), 141–144.
[15] O.R. Ibon, B. Garcia and Z.M. Amaia, New approach for oesophageal speech enhancement, 5th International Symposium on I/V Communications and Mobile Network (ISVC) **5** (2010), 225–228.
[16] J. Makhoul, Linear prediction: A tutorial review, Proceedings of the IEEE **63** (1975), 561–580.
[17] R.H. Ali and S.B. Jebara, Esophageal speech enhancement using excitation source synthesis and formant structure modification, Signal Processing for Image Enhancement and Multimedia Processing (SITIS) **31** (2006), 615–624.
[18] Y.Y. Qi, Replacing tracheoesophageal voicing source using lpc synthesis, Acoustical Society of America **5** (1990), 1228–1235.
[19] M. Alfredo, P.M. Hector, T. Jorge and O. Patricia, Analysis and recognition of esophageal speech, Symposium on Signal Processing and Information Technology **5** (2006), 101–106.
[20] Y.Y. Q., W. Bernd and B. Ning, Enhancement of female esophageal and tracheoesophageal speech, Acoustical Society of America **98** (1995), 2461–2465.
[21] K. Matsui and N. Hara, Enhancement of esophageal speech using formant synthesis, Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing **1** (1999), 81–84.
[22] R. Ishaq and B.G. Zapirain, Adaptive gain equalizer for improvement of esophageal speech, IEEE International Symposium on Signal Processing and Information Technology **1** (2012), 153–157.
[23] H. Doi, K. Nakamura, T. Toda, H. Saruwatari and K. Shikano, Statistical approach to enhancing esophageal speech based on gaussian mixture models, 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP) **10** (2010), 4250–4253.
[24] Y. Kamp and Willems. L. F. Ma, C, Robust single selection for linear prediction analysis of voiced speech, Speech Communication **2** (1983), 69–81.
[25] M.D.N. Westerlund and I. Claesson, Adaptive gain equalizer for speech enhancement, research report, Blekinge Institute of Technology, Karlskrona, 2002.
[26] C.P. Clark, Effective coherent modulation filtering and interpolation of long gaps in acoustic signals, M.S. Dissertation, University of Washington, 2008.
[27] T.K.J. Pohjalainen, R. Saeidi and P. Alku, Extended weighted linear prediction (xlp) analysis of speech and its application to speaker verification in adverse conditions, Interspeech, 2010, 1477–1480.
[28] W.-R. Wu, P.-C. Chen, H.-T. Chang and C.-H. Kuo, Frame-based subband Kalman filtering for speech enhancement, ICSP' 98 International Conference on Signal Processing Proceedings **1** (1998), 682–685.
[29] Pascal Clark Les Atlas and Steven Schimmel. Modulation toolbox version 2.1 for matlab, http://isdl.ee.washington.edu/projects/modulationtoolbox/, University of Washington, September 2010.
[30] R.E. Hillman and Y.Y. Qi, Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals, Acoustic Society of America **102** (1997), 537–543.
[31] A.A. Voicesauce, A program for voice analysis (2012), available at: http://www.seas.ucla.edu/spapl/voicesauce/, May 2014.
[32] I.M.-K.H. Kawahara and A.D. Cheveigne, Restructuring speech representation using a pitchadative time-frequency smoothing and an instantaneous-frequency-based f0 extraction, Speech Communication **27** (1999), 187–207.

Esophageal Speech Enhancement using Modified Voicing Source

Rizwan Ishaq and Begoña García Zapirain
DeustoTech-LIFE, University of Deusto, Spain
Email: rizwanishaq@deusto.es and mbgarciazapi@deusto.es

***Abstract*—This paper presented the modification to Esophageal Speech (ES) enhancement using Adaptive Gain Equalizer (AGE) for modifying the voicing source. However, the voicing source used previously with AGE, obtained using conventional Linear Predication (LP) vocal tract transfer function (AGE-LP), has produced low quality speech due to sensitivity to background noise. The better quality ES can be obtained by estimating voicing source through Iterative Adaptive Inverse Filtering utilized Weighted Linear Prediction (WLP) vocal tract transfer function (AGE-IAIF). The system performance evaluated through Harmonic to Noise Ratio (HNR), and system has shown 3 dB enhancement by AGE-IAIF over previously enhancement method AGE-LP.**

***Index Terms*—Filter bank, Iterative adaptive inverse filtering, Esophageal speech, Linear predictive coding, adaptive gain equalizer**

I. Introduction

The treatment of laryngeal cancer in advanced stages needs removal of vocal folds, which resulted in no voicing source for speech production. The alternative voicing source should be used for speech production, i.e. using esophagus or external devices. There are two methods uses esophagus, named Esophageal Speech (ES) and Tracheo-Esophageal Speech (TES), but with different air source. The other method Electrolarynx uses the external devices for production of voicing source. The air source for ES comes, by inhaling air into lower part of esophagus and then exhaled which vibrates the walls of esophagus and provides voicing source. The irregular vibration of voicing source in ES effects the quality of produced speech. Therefore speech enhancement methods needed for improving the quality of ES.

In literature, Linear Prediction (LP) decomposition of speech into source and filter components is used for enhancing the quality of ES and TES [1], [2]. The source component of speech signal replaced by LF source signal model for enhancing the ES [3]. The better quality ES is obtained modifying the formants of ES [4]–[7]. Besides LP source filter decomposition, other methods also used for enhancement purposes such as statistical methods [8], pole modification, Kalman filtering [9]–[15], and Adaptive Gain Equalizer (AGE) in modulation domain etc [16].

The paper has presented the ES enhancement method, considering the decomposition of ES into source filter components using Iterative Adaptive Inverse Filtering (IAIF) instead of LP decomposition. The AGE is used to modified the source of ES. The system is compared with [16], where source signal has been obtained using LP decomposition. The performance is measured through Harmonic to Noise Ratio (HNR). The paper is organized as follows, Sec. II discussed the system model components, such as IAIF decomposition, AGE etc. Sec. III discussed the experimental parameters, with results in Sec. IV and Conclusion in Sec. V.

II. System Model

The figure 1 shows the system model, with its components:

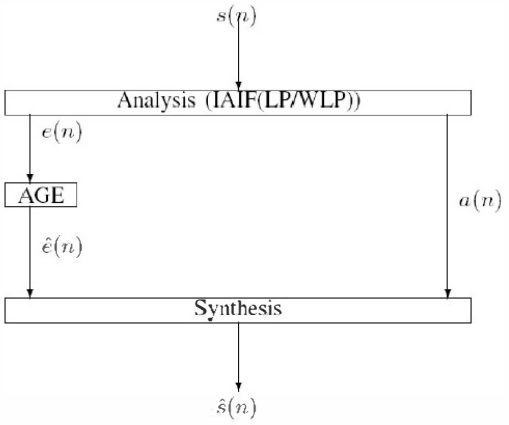

Fig. 1. Proposed System

A. Analysis (Iterative Adaptive Inverse Filtering (IAIF)

The Fant's theory of source-filter considered voiced speech signal as output of linear time-invariant vocal tract filter, excited by quasi periodic impulses [17]. Mathematically,

$$s(n) = e(n) * v(n) * r(n) \tag{1}$$

where $s(n)$ is speech signal, $e(n)$ source or excitation signal, $v(n)$ the vocal tract transfer function, and $r(n)$ is the lip radiation. The lip radiation $r(n)$ effect can be canceled, by approximating and modeling, as a derivative, and single zero FIR filter,

$$r(n) = 1 - \alpha r(n-1), \quad 0.95 < \alpha < 0.99 \tag{2}$$

In z-transform, after canceling lip radiation, we have:

$$S(z) = E(z)V(z) \tag{3}$$

978-1-4799-4796-6/13/$31.00 © 2013 IEEE

where $S(z)$, $E(z)$ and $V(z)$ are z-transform of $s(n)$, $e(n)$ and $v(n)$ respectively. The inversing filtering of speech signal $S(z)$ is used to obtained source signal $E(z)$:

$$E(z) = S(z)/V(z) \tag{4}$$

The transfer function $V(z)$ is compulsory component for inverse filtering, which can be estimated either by conventional LP or by IAIF source-filter decomposition. The IAIF uses Weighted Linear Prediction (WLP) for $V(z)$ as shown in Fig. 2. The Fig. 2 showed the complete IAIF system, where glottal transfer function $G(z)$ of order 1 for reducing the glottal source effect for optimal vocal tract transfer function $V(z)$. The optimal vocal tract transfer function $V(z)$ of order p (WLP) used to estimates source signal $e(n)$ through inverse filtering. In literature, $V(z)$ and $E(z)$ can be estimated using minimum-maximum phase decomposition [18], [19], IAIF [20] and LP decomposition [17] etc. This paper considered the IAIF decomposition (Weighted Linear Prediction (WLP)), and Linear Prediction (LP) decomposition [17], for source signal.

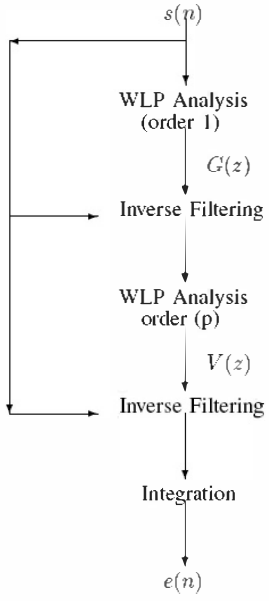

Fig. 2. Block diagram of IAIF(WLP) [20], [21]

1) Conventional Linear Prediction: The linear prediction model assumed, that speech signal can be estimated by a linear combination of p previous samples, mathematically:

$$\hat{s}(n) = \sum_{k=1}^{p} a_k s(n-k) \tag{5}$$

where $\hat{s}(n)$ is estimated speech signal at n instant, a_k prediction coefficients, and p order of prediction. The prediction error E is used to estimated the a_k coefficients by minimizing its sum of squares [17], [22],

$$E = \sum_{n} (s(n) - \sum_{k=1}^{p} a_k s(n-k))^2 \tag{6}$$

$$\frac{\partial E}{\partial a_k} = 0 \Longrightarrow \sum_{k=1}^{p} a_k \sum_{n} s(n-k)s(n-j) = \sum_{n} s(n)s(n-j) \tag{7}$$

Solving the above equation for a_k $(1 \leq j \geq p)$, the vocal tract transfer function can be estimated as:

$$V(z) = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-}k} \tag{8}$$

which subsequently uses equation.(4) for source signal $e(n)$ [22].

2) Weighted Linear Prediction (WLP): The WLP is a modification to linear prediction, introducing temporal weighting function to the square prediction error according to following mathematical relation [23], [24]:

$$E = \sum_{n} (s(n) - \sum_{k=1}^{p} a_k s(n-k))^2 W(n) \tag{9}$$

where $W(n)$ is the weighting function and given as short-time energy function [25]:

$$W(n) = \sum_{k=1}^{M} s^2(n-k) \tag{10}$$

where M is number of samples for calculating energy (M=16 used in this experiment). Solving for prediction error, prediction coefficients a_k can be obtained as [23], [24]:

$$\frac{\partial E}{\partial a_k} = 0 \tag{11}$$

$$\sum_{k=1}^{p} a_k \sum_{n} W(n)s(n-k)s(n-i) = \sum_{n} W(n)s(n)s(n-i) \tag{12}$$

The obtained predication coefficients $a_k (1 \leq i \geq p)$ are used for vocal tract transfer function $V(z)$(equation. (8)), which gives source signal $e(n)$ according to equation.(4).

B. Adaptive Gain Equalizer (AGE)

The AGE is robust and standalone speech enhancement method which enhanced periodic part of speech signal by raising Signal to Noise Ratio (SNR) of sub-bands shown in figure. 3 [26]–[30]. The modified source signal $\hat{e}(n)$ is obtained by applying AGE process to source signal $e(n)$.

1) Filterbank(Analysis): The source signal $e(n)$ passed through uniformaly-spaced subbands in a modified Short-Time Fourier Transform (STFT) filterbank of K bandpass filters [31], each having the impulse response of $h_k(n)$ [26]–[30][1].

$$e_k(n) = h_k(n) * e(n) \tag{13}$$

The $*$ is convolution operator. Each subband signal $e_k(n)$ is enhanced by weighting it, according to SNR gain function $w_k(n)$ of that sub-band,

$$\hat{e}_k(n) = w_k(n) e_k(n) \tag{14}$$

[1]This section based on [32]–[35]

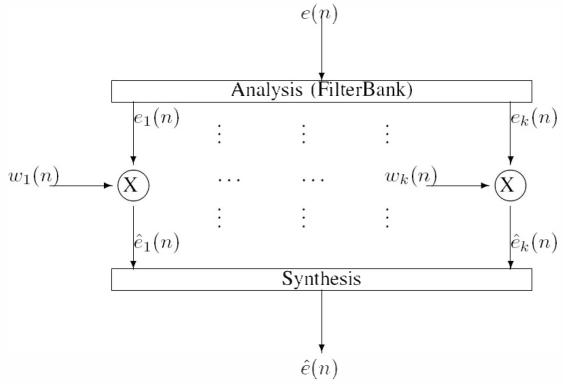

Fig. 3. Block diagram AGE

2) Weighting function: The $w_k(n)$ is the ratio of short term average and long term average of each sub-band signal, according to following mathematical relation:

$$w_k(n) = min\left\{\left(\frac{\mu_{k,f}(n)}{L_{opt}.\mu_{k,s}(n)}\right)^{p_k}, L_k\right\} \tag{15}$$

where $\mu_{k,f}$ is short term or fast average, $\mu_{k,s}$ long term or slow average, L_{opt} the optimal suppression level to control $w_k(n)$, L_k the limiting threshold, and p_k the gain rise exponent component. The short term and long term averages are given:

$$\mu_{k,f}(n) = \alpha_k \mu_{k,f}(n-1) + (1-\alpha_k)|e_k(n)| \tag{16}$$

where $\alpha_k = \frac{1}{f_{k,s}T_{k,a}}$ is forgetting factor constant, $f_{k,s}$ and $T_{k,a}$ are sampling frequency and time constant for sub-bands respectively.

$$\mu_{k,s}(n) = \begin{cases} \mu_{k,f}(n) & \text{if } \mu_{k,f}(n) \leq \mu_{k,s}(n-1) \\ (1+\beta_k)(\mu_{k,s}(n-1)) & \text{otherwise} \end{cases} \tag{17}$$

where $\beta_k = \frac{1}{f_{k,s}T_{k,b}}$ is a positive constant control the noise level utilizing time constant $T_{k,b}$

3) Synthesis: The enhanced version of source/excitation signal $\hat{e}(n)$ is obtained by summing all the sub-band signals according to following relation:

$$\hat{e}(n) = \sum_{k=0}^{K-1} \hat{e}_k(n) \tag{18}$$

C. Synthesis

The enhanced version of source signal is used to excite the vocal tract transfer function $V(z)$ (utilizing prediction coefficients $(a_k(n))$ for enhanced version of ES signal,

$$\hat{S}(z) = \hat{E}(z)V(z) \tag{19}$$

The time domain enhanced signal $\hat{s}$ is obtained by inverse z-transform of $\hat{S}(z)$.

III. Experiment Parameters

The ES Spanish vowels \a\, \e\, \i\, \o\, \u\(20 utterance for each vowel) are used for this experiment, recorded from rehabilitation center. The sampling frequency for the recording was 44100 Hz and down-sampled to 16000 Hz for computational efficiency. The 6 persons who have very good quality of ES, recorded their voices. The prediction order p is 12 for for both LP and WLP, the segment size and overlap size are set 30ms and 15 ms respectively for estimation of voicing source. The AGE system which modified the voicing source $e(n)$ has 16 channel filterbank, with decimation factor of 4. The other parameter used in AGE are shown in table I.

TABLE I
Parameter values for Adaptive Gain Equalizer

Parameter	Value
$T_{k,a}$	30 msec
$T_{k,b}$	3 msec
L_{opt}	$0 \rightarrow 20$
L_k	30 dB
p_k	1

IV. Results

A. Harmonic to Noise Ratio (HNR)

The Harmonic to Noise Ratio (HNR), is one of the objective acoustic measure for speech signal engineers to quantify noise level in the signal, calculated as the ratio of harmonic energy and noise energy components of speech [36]. The estimation of HNR assumed that speech signal consists of periodic and additive noise components. The methods [36], [37] considered, the original speech signal $f(t)$, as the concatenation of the waves $f_k(\tau)(k$ signal period) from each pitch period, where τ is the number of pitch period. Averaging $f_k(\tau)$ for large number n gives us the better estimation of harmonic components,

$$f_A(\tau) = \sum_{k=1}^{n} \frac{f_k(\tau)}{n} \tag{20}$$

The energy of the harmonic components is calculated as,

$$H = n\sum_{\tau=1}^{T} f_A^2(\tau) \tag{21}$$

The noise components N energy is defined as,

$$N = \sum_{i=1}^{n}\sum_{\tau=0}^{T_i}[f_i - f_A(\tau)]^2 \tag{22}$$

where T_i is the period length in samples. The HNR is given by as,

$$HNR = 10log_{10}\frac{H}{N} \tag{23}$$

The VoiceSauce [38], freely available speech analysis software, used to calculate the HNR. The parameters used for HNR calculation are, 25 milliseconds windows size with 1 millisecond overlap, fundamental frequency is constrained between 50 Hz to 120 Hz (ES fundamental frequency falls in

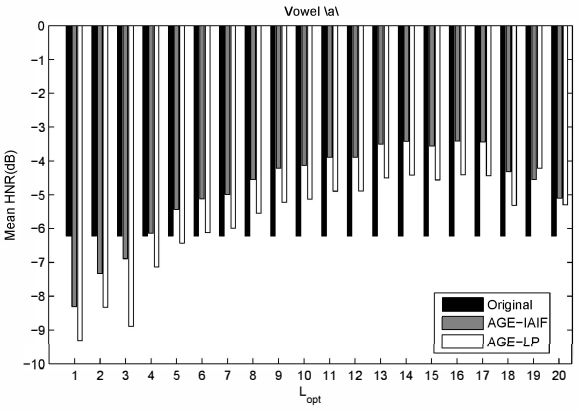

Fig. 4. Mean Harmonic to Noise Ratio (HNR) for vowel a

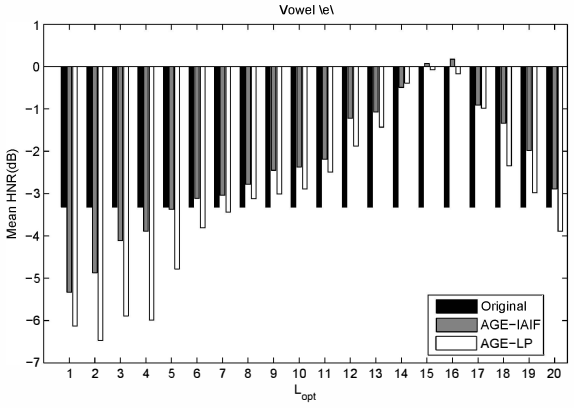

Fig. 5. Mean Harmonic to Noise Ratio (HNR) for vowel e

this range). The HNR obtained for different values of L_{opt} (optimal suppression gain controller). The figures 4 and 5 has shown improvement of 3 dB for AGE-IAIF in comparison to AGE -LP [16], particularly rising trend observed between L_{opt} optimized values 8 - 16. The starting and ending values of L_{opt} has decreasing effect in HNR for all the vowels, therefore system should be tuned between L_{opt} values of 8-16 for optimal results. On global scale, system has provided better results for vowels \a\, \e\, \i\but vowels \o\and \u\has not produced optimal results, as shown in figures 7 and 8, with only 1 dB improvement.

V. Conclusion

The paper has discussed the performance of ES enhancement by modifying voicing source signal through AGE system. The voicing source has been obtained by applying IAIF to speech signal, by estimating vocal tract transfer function through Weighted Linear Prediction (WLP). The AGE modified the voicing source and compared with previously available system, where conventional LP vocal tract transfer function estimation used for voicing source estimation [16] Using the Harmonic to Noise Ratio (HNR) criterion, its has been shown that AGE with IAIF (AGE-IAIF) provided better results in comparison to AGE with conventional LP (AGE-LP) [16]. In future system can be further improved by implementing vocal tract formant enhancement as well the voicing source with other prediction methods such as stabilized WLP, extended LP etc.

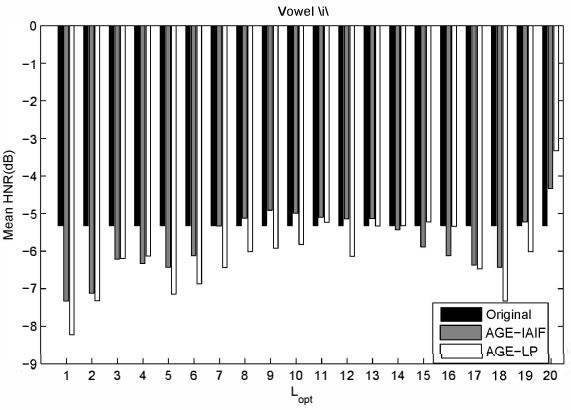

Fig. 6. Mean Harmonic to Noise Ratio (HNR) for vowel i

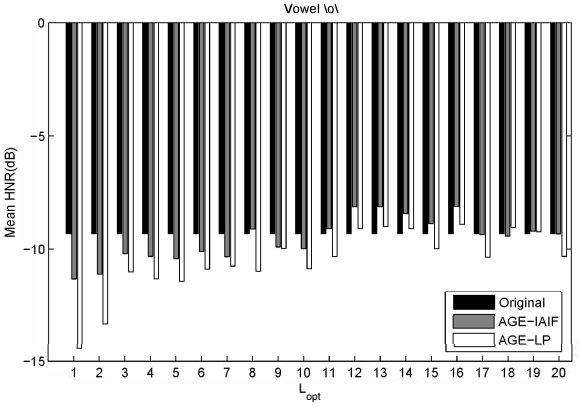

Fig. 7. Mean Harmonic to Noise Ratio (HNR) for vowel o

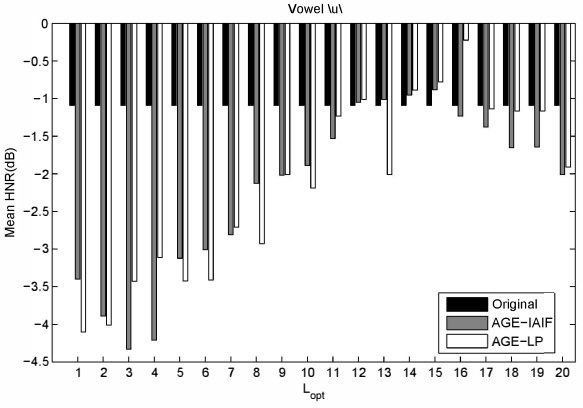

Fig. 8. Mean Harmonic to Noise Ratio (HNR) for vowel u

ACKNOWLEDGMENT

This research was partially granted by Deiker of Deusto University and Department of Education and Researcher of Basque government under the support of Duestotech eVida group of Deusto university project.

REFERENCES

[1] Q. Yingyong, W. Bernd, and B. Ning, "Enhancement of female esophageal and tracheoesophageal speech," *Acoustical Society of America*, vol. 3, pp. 1228–1235, 1990.

[2] Q. Yingyong, "Replacing tracheoesophageal voicing source and lpc synthesis," *Acoustical Society of America*, vol. 5, pp. 2461–2465, 1995.

[3] R. Sirichokswad, P. Boonpramuk, N. Kasemkosin, P. Chanyagorn, W. Charoensuk, and H. H. Szu, "Improvement of esophageal speech using lpc and lf model," *Internation Conf. on Biomedical and Pharamaceutical Engineering 2006*, pp. 405–408, 2006.

[4] M. Kenji and H. Noriyo, "Enhancement of esophageal speech using formant synthesis," *Acoustics, Speech and Signal Processing, International conf.*, pp. 81–85, 1999.

[5] R. H. Ali and S. B. Jebara, "Esophageal speech enhancement using excitation source synthesis and formant structure modification," *IEEE*, 2005.

[6] R. A. Prosek and L. L. Vreeland, "The intelligibility of time domain edited esophageal speech," *American Speech Language Hearing Association*, vol. 44, pp. 525–534, 2001.

[7] A. Loscos and J. Bonada, "Esophageal voice enhancement by modeling radiated pulses in frequency domain," *Audio Engineering Society*, 2006.

[8] H. Doi, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Statistical approach to enhancing esophageal speech based on gaussian mixture models," *Acoustics Speech and Signal Processing(ICASSP)*, pp. 4250–4253, 2010.

[9] B. Garcia and A. Mendez, "Oesophageal speech enhancement using poles stablization and kalman filtering," *ICASSP*, pp. 1597–1600, 2008.

[10] O. Ibon, B. Garcia, and Z. M. Amaia, "New approach for oesophageal speech enhancement," *10th International conference, ISSPA*, vol. 5, pp. 225–228, 2010.

[11] B. Garcia and J. Vincente, "Software for measuring and improving esophageal voice," *Internation Conference on Digital Audio Effects(DAFx04), Italy*, vol. 5, pp. 303–306, 2004.

[12] B. Garcia, I. Ruiz, A. Mendez, and M. Mendezona, "Oesophageal voice acoustic parameterization by means of optimum shimmer calculation," *WSEAS Trasactions on Systems*, pp. 489–499, 2008.

[13] B. Garcia, I. Ruiz, J. Vicente, and A. Alonso, "Formants measurement for esophageal speech using wavelet with band and resoultion adjustment," *IEEE Symposium on Signal Processing and Information Technology*, pp. 320–325, 2006.

[14] B. Garcia, I. Ruiz, A. Mendez, and M. Mendezona, "Objective characterization of oesophageal voice supporting medical diagnosis rehabilitation and monitoring," *Computers in Biology and Medicine, Elsevier*, pp. 97–105, 2009.

[15] B. Garcia, J. Vicente, A. Alonso, and E. Loyo, "Esophageal voices: glottal flow restoration," *Acoustics, Speech and Signal Processing 2005(ICASSP 05)*, pp. 141–144, 2005.

[16] R. Ishaq and B. G. Zapirain, "Adaptive gain equalizer for improvement of esophageal speech," in *IEEE International Symposium on Signal Processing and Information Technology*, 2012.

[17] G. Fant, "Acoustic theory of speech production." Mouton, The Hauge, 1960.

[18] D. B. D. A. C. D. T. Bozkurt, B., "Zeros of z-transform representation with application to source-filter separation in speech," in *IEEE Signal Processing Letters*, vol. 12, no. 4, pp. 344–347.

[19] B. D. T. Drugman, T. Bozkurt, "Complex cepstrum-based decomposition of speech for glottal source estimation," in *Proc. Interspeech*, 2009, pp. 116–119.

[20] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," in *Speech communication*, vol. 11, no. 2, 1992, pp. 109–118.

[21] M. V. A. Suni, T. Raitio and P. Alku, "The glottalhmm entery for blizzard challenge 2011: Utilizing source unit selection in hmm-based speech synthesis for improved excitation generation," in *in Blizzard Challenge 2011, Workshop, Florence, Italy*, 2011.

[22] S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," 1971.

[23] T. K. Jouni Pohjalainen, Rahim Saeidi and P. Alku, "Extended weighted linear prediction (xlp) analysis of speech and its application to speaker verification in adverse conditions," in *INTERSPEECH, Annual Conference of the International Speech Communication Association*, 2010.

[24] B. T. Magi, C. Pohjalainen and P. Alku, "Stablized weighted linear prediction," in *Speech Communcation*, vol. 5, no. 51, 2009, pp. 401–411.

[25] K. Y. Ma, C. and L. F. Willems, "Robust signal selection for linear prrediction analysis of voiced speech," *Speech Communication*, vol. 12, pp. 69–81, 1993.

[26] M. D. Nils Westerlund and I. Claesson, "Adaptive gain equalizer for speech enhancement," Blekinge Institute of Technology, Karlskrona, Research Report, 2002.

[27] R. Ishaq and B. G. Zapirain, "Adaptive gain equalizer for improvement of esophageal speech," *IEEE International Symoposium on Signal Processing and Information Technology (ISSPIT 2012)*, 2012.

[28] N. Westerlund, M. Dahl, and I. Claesson, "Speech enhancement using an adaptive gain equalizer with frequency dependent parameter settings," *VTC04*, 2004.

[29] ——, "Speech enhancement for personal communication using an adaptive gain equalizer," *Elsevier Signal Processing.*, vol. 85, pp. 1089–1101, 2005.

[30] ——, "Real-time implementation of an adaptive gain equalizer for speech enhancement purposes," *WSEAS.*, 2003.

[31] S. M. Schimmel, "Theory of modulation frequency analysis and modulation with application to hearing devices," Ph.D Thesis, University of Washington, 2007.

[32] R. Ishaq, "Adaptive Gain Equalizer and Modulation Frequency Domain for Noise Reduction," Master Thesis, Blekinge Institute of Technology, Karlskrona, Sweden.

[33] S. Muhammad, I. Rizwan, S. Benny, G. Nedelko, L. Benny, and C. Ingvar, "Modulation domain adaptive gain equalizer for speech enhancement," *IASTED International Conference on Signal and Image Processing and Applications*, 2011.

[34] B. et al, "An improved adaptive gain equalizer for noise reduction with low speech distortion," *EURASIP Journal on Audio Speech and Music Processing*, 2011.

[35] N. Westerlund, M. Dahl, and I. Claesson, "Speech enhancement for personal communication using an adaptive gain equalizer," *EURASIP Journal on Audio Speech and Music Processing*, pp. 1089–1101, 2005.

[36] E. Yumoto and W. J. Gould, "Harmonics-to-noise ratio as an index of the degree of hoarseness," *Acoustic Society of America*, vol. 71, pp. 1544–1550, 1982.

[37] R. Sousa and A. Ferriera, "Evaluation of existing harmonics-to-noise ratio methods for voice assessment," in *Signal Processing Algorithms, Architectures, Arrangements, and Applications (SPA)*, 2008, pp. 73–78.

[38] A. Alwan. (2012, Feb.) Voicesauce: A program for voice analysis @ONLINE. [Online]. Available: http://www.ee.ucla.edu/ spapl/voicesauce/

Subband Modulator Kalman Filtering for Single Channel Speech Enhancement

Rizwan Ishaq*, Begoña García Zapirain*, Muhammad Shahid** and Benny Lövström**

*University of Deusto, Bilbao, Spain
**School. of Elec. Engineering, Blekinge Institute of Technology, Karlskrona, Sweden

***Abstract*— *This paper presents a single channel speech enhancement technique based on sub-band modulator Kalman filtering for laryngeal (normal) and alaryngeal (Esophageal speech) speech signals. The noisy speech signal is decomposed into sub-bands and subsequently each sub-band is demodulated into its modulator and carrier components. Kalman filter is applied to modulators of all sub-bands without altering the carriers. Performance of the proposed system has been validated by Mean Opinion Score (MOS) for laryngeal and Harmonic to Noise Ratio (HNR) for alaryngeal speech. An improvement of 20% has been observed in MOS over sub-band Kalman filtering for laryngeal speech, while 3 to 4 dB enhancement in HNR has been observed for alaryngeal speech over the full-band Kalman filtering.*

Keywords*— *Kalman filter, Autoregressive, speech enhancement

I. Introduction

Speech enhancement is an important branch of speech signal processing that aims at suppression of noise to make a speech signal more intelligible. An enhanced version of a speech signal is useful for speech recognition applications, mobile communication and coding etc. There has been many algorithms proposed for speech enhancement including but not limited to spectral subtraction [1], [2], Wiener filtering [3], adaptive gain equalizer [4], [5], [6], [7] and Kalman filtering [8], [9].

Kalman filtering is considered to be an optimal speech enhancement algorithm that relies on a Minimun Mean Square Error (MMSE) [10], [8] based method. The Kalman filtering based speech enhancement has several advantages over other speech enhancement methods, e.g. speech production model using Linear Predication (LP), inherited to Kalman filtering modeling. Kalman filter produces optimum results for non-stationary signals and do not need stationary condition like Wiener filtering [10].

The Kalman filter is used for single channel speech enhancement by Analysis-Modification-Synthesis (AMS) frame work, where noisy speech signal is segmented into frames using short time Fourier transform (STFT), then a modification of amplitude of STFT is applied using Kalman filtering followed by inverse STFT and synthesis for enhanced speech signal [11]. Paliwal introduced the Kalman filtering for speech enhancement [8]. Further modification to Kalman filtering has been observed using the EM algorithm for autoregressive (AR) estimation for Kalman filtering [12], [13], [14]. The enhancement for colored noise corrupted speech has also been investigated in [15] using Kalman filter. The most important and less complex modification done by sub-band based Kalman filtering for speech enhancement is by dividing the speech signal into a number of sub-bands followed by Kalman filtering of each sub-band [16], [17].

The Esophageal (E) speech is one type of alaryngeal speeches used for speech production after laryngeal cancer treatment, where larynx has been removed and normal speech in no more possible. The E speech has low quality due to irregular vibration of Paryngo-esophageal (PE) segments, and enhancement of E speech has been extensively treated by LPC analysis/synthesis [18], [19], [20], [21], [22], statistical methods [23], [24], [25] and detailed analysis of E speech by our group can be consulted from [26], [27], [28], [29], [30], [31], [32]. The Kalman filter has been used for enhancement of E speech along with pole stabilization and, improvement observed over LPC analysis/synthesis framework [33], [34].

Recent research has used the approach to model speech signals as the combination of low and high frequency components, called modulators and carriers respectively. The modulators (low frequency) are considered to be most important for speech intelligibility, i.e. if speech modulators are replaced by a constant value, while preserving carriers, unintelligible speech is obtained, in comparison to the case of preserving modulators and replacing carriers with constant value retains the intelligibility of speech [35]. Mathematically,

$$x(n) = m(n)c(n) \tag{1}$$

where $m(n)$ and $c(n)$ are modulators and carriers respectively. A trend has been observed in recent years that speech enhancement by modifying modulators of speech signal is done using different techniques. Results justify the use of modulator filtering, e.g. convex optimization, and center of gravity (CoG) demodulation, used to enhance speech signals [36], [37].

This paper introduces a modification to sub-band based Kalman filter based speech enhancement [16], by decomposing sub-bands into its modulators and carriers components. The Kalman filter is applied to modulators of sub-bands instead of sub-bands directly. Performance of the system has been validated by Mean Opinion Score (MOS) and spectrogram for laryngeal (normal) speech by comparing it to sub-band Kalman filtering [16], and Harmonic to Noise Ratio (HNR) used for alaryngeal (E speech) by comparing it with full-band Kalman filtering E speech enhancement [33], [34]. The next sections introduce system components followed by results and conclusion.

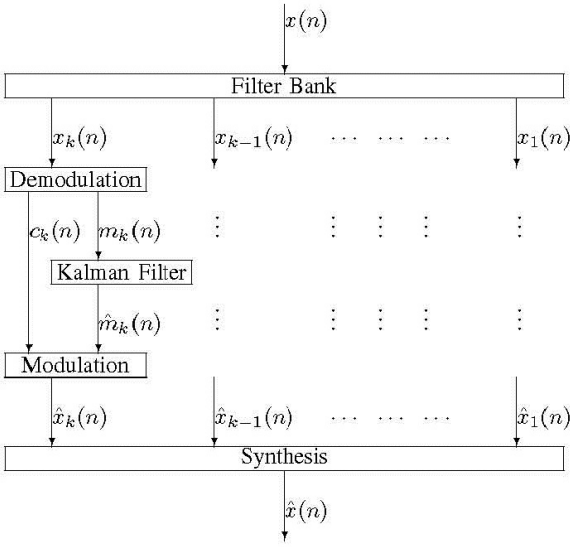

Fig. 1. Sub-band Modulator Kalman Filtering Based Speech Enhancement

II. System Design

Fig. 1 shows the proposed system used for the enhancement of noisy speech signal $x(n)$.

A K bands band-pass filter is used to divide the input speech signal $x(n)$ into sub-bands according to:

$$x_k(n) = h_k(n) * x(n) \tag{2}$$

where $h_k(n)$ is impulse response of the kth sub-band filter and $*$ is convolution operator. Each sub-band is demodulated into modulator $m_k(n)$ and carrier $c_k(n)$ coherently according to CoG demodulation (Section III-A).

$$x_k(n) = m_k(n)c_k(n) \tag{3}$$

Sub-band modulators are modified by Kalman filtering (Section IV), given by:

$$\hat{x}_k(n) = \hat{m}_k(n)c_k(n) \tag{4}$$

where $\hat{m}_k(n)$ is modified modulator for sub-band k. The final enhanced signal is obtained by adding all the modified sub-bands according to the synthesis equation:

$$\hat{x}(n) = \sum_{k=1}^{K} \hat{x}_k(n) \tag{5}$$

III. Demodulation

Natural signals such as speech can be represented by the corresponding high frequency and low frequency components, called carriers and modulators respectively [35], [38], [39], [40]. The speech signal can be represented (in modulators and carriers sense) by equation (1). The decomposition of speech signal into $m(n)$ and $c(n)$ can be acquired coherently or non-coherently [35], [39], [40]. The non-coherent demodulation estimates the modulators and carriers independent of each other, while in coherent demodulation carriers are estimated first and then modulators are estimated based on the equation (1). In this paper, coherent demodulation has been used because of its advantages over the non-coherent and in the present case, carrier estimation is done using spectral center of gravity [41], [35], [42].

A. Spectral Center of Gravity Carrier Estimation

The demodulation framework works on sub-bands, the filter bank divides the speech signal into sub-bands, demodulation process decomposes each sub-band into its carrier and modulator components.

1) Sub-band Instantaneous Frequency: The first step in calculating the carrier is to detect the instantaneous frequency $\omega_k(n)$ of each sub-band. The center of gravity approach estimates the $\omega_k(n)$ as the average frequency of instantaneous spectrum of $x_k(n)$ [41], [35]. The instantaneous spectrum of x_k is calculated according to:

$$S_k(\omega, n) = \sum_p g(p)x_k(n+p)e^{-j\omega p} \tag{6}$$

where $g(p)$ is a window function (hamming window of length 128 is used for this experiment). Center of Gravity (CoG) estimation of $\omega_k(n)$ is given by:

$$\omega_k(n) = \frac{\int_{-\pi}^{\pi} \omega |S_k(\omega, n)|^2 d\omega}{\int_{-\pi}^{\pi} |S_k(\omega, n)|^2 d\omega} \tag{7}$$

The phase $\phi_k(n)$ is obtained by the following equation:

$$\phi_k(n) = \sum_{p=0}^{n} \omega_k(p) \tag{8}$$

2) Carrier estimation: Carrier $c_k(n)$ obtained by exponentiating $\phi_k(n)$:

$$c_k(n) = exp[j\phi_k(n)] \tag{9}$$

The carrier estimation for sub-band k gives the related modulator as:

$$m_k(n) = x_k(n)/c_k(n) = x_k(n)c_k^*(n) \tag{10}$$

IV. Subband Modulator Kalman Filtering

It is considered that modulators of speech signal can be represented by an autoregressive (AR) process, i.e. output of an all-pole system excited by white Gaussian noise and represented by a difference equation:

$$m_k(n) = \sum_{j=1}^{p} a_{k,j} m_k(n-j) + w_k(n) \tag{11}$$

where $a_{k,j}(n)$, p and $w_k(n)$ are Linear Predication Coefficients (LPC), order of AR process and input white Gaussian noise(with zero mean and variance $\sigma_{k,w}^2$) respectively for the

kth sub-band modulator $m_k(n)$. The observed noisy modulator for sub-band k is given by $s_k(n)$ as:

$$s_k(n) = m_k(n) + v_k(n) \tag{12}$$

where $v_k(n)$ is white Gaussian additive observation or measurement noise with zero mean and variance $\sigma^2_{k,v}$ for sub-band k. The equations given above can be given in the state space representation as:

$$m_k(n) = F_k m_k(n-1) + g w_k(n) \tag{13}$$

$$s_k(n) = H^T m_k(n) + v_k(n) \tag{14}$$

where $m_k(n) = [m_k(n-p+1) m_k(n-p+2) \cdots m_k(n)]$.

$$F_k = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ -a_{k,p} & -a_{k,p-1} & -ak,p-2 & \dots & -ak,1 \end{bmatrix} \tag{15}$$

$$g^T = H^T = [0, 0, \dots, 1] \tag{16}$$

The Kalman filter provides the estimate of $m_k(n)$, providing observation $s_k(1), s_k(2), ...s_k(n)$ [15] as:

$$\hat{m}_k(n) = F_k \hat{m}_k(n-1) + K_k(n)[s_k(n) - H^T F_k \hat{m}_k(n-1)] \tag{17}$$

$$K_k(n) = P_k(n|n-1)H[R_k + H^T P_k(n|n-1)H]^{-1} \tag{18}$$

$$P_k(n|n-1) = F_k P_k(n-1|n-1) F_k^T + g Q_k g^T \tag{19}$$

$$P_k(n) = [I - K_k(n) h^T] P_k(n|n-1) \tag{20}$$

where $K_k(n)$ is Kalman gain, $P_k(n|n-1)$ is a priori error covariance matrix and $P_k(n)$ is error covariance matrix, R_k and Q_k are measurement noise covariance matrix and input noise covariance matrix respectively for sub-band k. The system is initialized using the noisy modulator:

$$\hat{m}_k(0) = m_{k,0} = [s_k(1), s_k(2), \dots, s_k(p)] \tag{21}$$

$$P_k(0|0) = P_{k,0} = diag[R_k, R_k, \dots, R_k] \tag{22}$$

At time instant n estimated sample is given by following relationship:

$$\hat{m}_k(n) = H^T \hat{m}_k(n) \tag{23}$$

A. Parameter Estimation

The estimation of LPC coefficients and noise variances for sub-band modulators is necessary for optimized results of Kalman filter. These parameters of each sub-band are calculated based on EM algorithm given in [12] and it is given below breifly:

- Noisy only segment from modulator of sub-band k is detected, and additive observation noise $\sigma^2_{k,v}$ is estimated.
- LPC parameters $a_{k,n}$ and variance $\sigma^2_{k,nmodulator}$ are calculated for noisy speech modulator.
- Input noise variance is estimated by $\sigma_{k,w} = \sigma_{k,nmodulator} - \sigma^2_{k,v}$

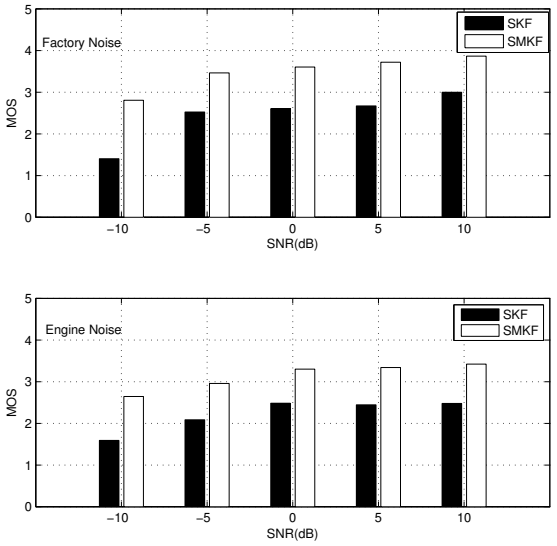

Fig. 2. Mean Opinion Score for Sub-band Kalman filter (SKF) and Sub-band Modulator Kalman Filter (SMKF).

- Kalman filter is implemented with noisy parameters, then enhanced version of modulator is used to estimate $a_{k,n}$, iterated until optimal estimate is obtained. In our work, the number of iteration are 3 as stated in [12].

V. Comparative Performance Analysis

A. Laryngeal Speech

Performance of the system has been tested using female speech signal sampled at 16 KHz, and corrupted by factory and engine noise signals with different Signal to Noise Ratio (SNR) (-10, -5, 0, 5, 10 dB). The number of filters in the filter bank effects the results, for this work, the number of filter used are 64 which gave better results in reducing residual noise. The Kalman filter uses the LPC order p of 10, and window size and step sizes are 30 and 15 millisecond respectively. This paper presents the comparison between systems based on MOS and spectrogram.

1) Mean Opinion Score (MOS): Fig. 2 shows the comparison of enhanced version speech signal with Sub-band Kalman Filtering (SKF) and Sub-band Modulator Kalman Filtering (SMKF) for MOS values. A maximum of 20% improvement can be observed and SMKF outperforms SKF for all SNR cases.

2) Spectrogram: Fig.3 and 4 show the spectrogram of speech signal corrupted by engine noise and factory noise at -10dB SNR. Although SMKF shows some loss in formants in upper frequencies but in comparison to SKF, there is less residual noise in enhanced speech signal. Significant improvement can be observed in factory noise corrupted speech signal.

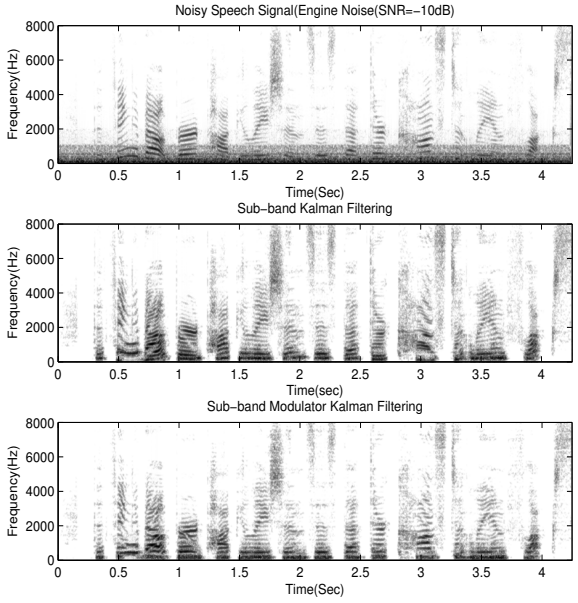

Fig. 3. Spectrogram of noisy and enhanced speech signals by systems(Engine Noise)

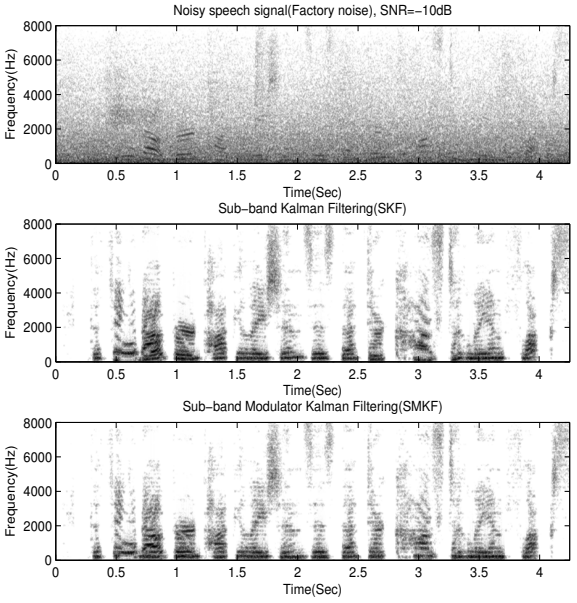

Fig. 4. Spectrogram of noisy and enhanced speech signals by systems(Factory Noise)

B. Alaryngeal Speech

The E speech vowels \a\, \e\,\i\, \o\, \u\, and \bodega\have been used for this experiment, which are recorded from alaryngeal speech rehabilitation center (6 persons) with the sampling frequency of 44100 Hz and down-sampled to 16000 Hz for computational efficiency.

1) Harmonic to Noise Ratio (HNR): VoiceSauce [43] was used to calculate HNR, with following settings, fundamental frequency range: 60 to 120 Hz (E speech fundamental frequency range is in between 60-120), frame length and overlap was set to 30 and 15 millisecond respectively and LPC order was set to 12. Fig. 5 shows the improvement of around 4 dB over the full-band Kalman filtering [33], [34], and 2 dB over sub-band Kalman filtering.

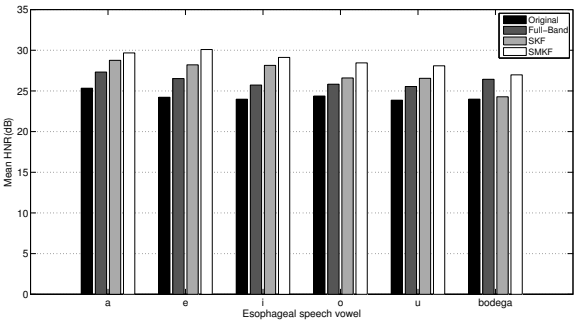

Fig. 5. Mean Harmonic to Noise Ratio (HNR)

VI. Conclusion

The modification to sub-band Kalman filtering by applying Kalman filter to modulators of sub-band by coherent decomposition has been successfully implemented for noisy laryngeal and alaryngeal speeches (E speech). Results thus obtained show improvement in speech enhancement while Kalman filtering is used in modulator domain in comparison to its traditional use. The improvement in MOS and spectrogram has shown the system capability of the proposed for reducing noise from noisy laryngeal speech, and HNR improvement has confirmed the system performance over the previous methods for E speech. The future work can be the utilization of other demodulation process, e.g. non-coherent demodulation and convex optimization demodulation [36], [44], [45].

References

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE trans. Accoust. Speech and Sig. Proc.*, vol. 27, no. 2, pp. 113–120, 1979.

[2] K. Paliwal, K. Wójcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech Commun.*, vol. 52, no. 5, pp. 450–475, May 2010.

[3] M. H. Hayes, *Statistical Digital Signal Processing and Modeling*, 1st ed. New York, NY, USA: John Wiley & Sons, Inc., 1996.

[4] M. Dahl and B. Sallberg, "Speech enhancement implementations in the digital, analog and hybrid domain," *Swedish System on Chip Conference*, 2005.

[5] N. Westerlund, M. Dahl, I. Claesson, B. Sallberg, and H. Akesson, "Analog circuit implementation for speech enhancement purposes," *Asilomar Conference on Circuits, Systems and Computers.*, 2004.

[6] M. Dahl, I. Claesson, B. Sallberg, and H. Akesson, "A mixed analog -digital hybrid for speech enhancement purposes," *ISCAS.*, 2005.

[7] N. Westerlund, M. Dahl, and I. Claesson, "Speech enhancement for

personal communication using an adaptive gain equalizer," *Elsevier Signal Processing.*, vol. 85, pp. 1089–1101, 2005.

[8] K. K. Paliwal and A. Basu, "A speech enhancemet method based on kalman filtering," *IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 12, pp. 177–180, 1987.

[9] B. K. J. D. Gibson and S. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Trans. , Signal Processing*, vol. 39, no. 8, pp. 1732–1742, 1991.

[10] S. So and K. K. Paliwal, "Supressing the influence of additive noise on the kalman gain for low residual noise speech enhancement," *Elsever, Speech communication 53*, vol. 53, pp. 355–378, 2011.

[11] ——, "Modulation-domain kalman filtering for single-channel speech enhancement," *Speech Commun.*, vol. 53, no. 6, pp. 818–829, Jul. 2011.

[12] D. Weixiu and P. Driessen, "Speech enhancement based on kalman filtering and em algorithm," *IEEE Pacific Rim Conf. on Communication, Computers and Signal Processing, 1991*, pp. 142–145, 1991.

[13] S. So and K. K. Paliwal, "A long state vector kalman filter for speech enhancement," in *Proceedings of the 9th Annual Conference of the International Speech Communication Association*, 2008, pp. 391–394.

[14] E. G. M. Gabrea and M. Najim, "A single microphone kalman filter-based noise cancellor," *IEEE Signal processing Letters*, vol. 6, no. 3, pp. 55–57, 1999.

[15] D. J. G. B. Koo and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Trans. on Signal Processing*, no. 8, pp. 1732–1742, 1991.

[16] W.-R. Wu and P.-C. Chen, "Subband kalman filtering for speech enhancement," *Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on*, vol. 45, no. 8, pp. 1072–1083, 1998.

[17] C. H. You, S. N. Koh, and S. Rahardja, "Subband kalman filtering incorporating masking properties for noisy speech signal," *Speech Commun*, vol. 49, no. 7-8, pp. 558–573, Jul 2007.

[18] M. Kenji and H. Noriyo, "Enhancement of esophageal speech using formant synthesis," *Acoustics, Speech and Signal Processing, International conf.*, pp. 81–85, 1999.

[19] M. Alfredo, P. Hector, T. Jorge, and O. Patricia, "Analysis and recognition of esophageal speech," *Symposium on Signal Processing and Information Technology*, vol. 5, pp. 101–106, 2006.

[20] Y. Qi, "Replacing tracheoesophageal voicing source using lpc synthesis," *Acoustical Society of America*, vol. 5, pp. 1228–1235, 1990.

[21] R. Sirichokswad, P. Boonpramuk, N. Kasemkosin, P. Chanyagorn, W. Charoensuk, and H. H. Szu, "Improvement of esophageal speech using lpc and lf model," *Internation Conf. on Biomedical and Pharamaceutical Engineering 2006*, pp. 405–408, 2006.

[22] Q. Yingyong, W. Bernd, and B. Ning, "Enhancement of female esophageal and tracheoesophageal speech," *Acoustical Society of America*, vol. 98(5, Pt1), pp. 2461–2465, 1995.

[23] K. T. T. S. H. S. K. Doi, H.; Nakamura, "Statistical approach to enhancing esophageal speech based on gaussian mixture models," *Acoustics Speech and Signal Processing(ICASSP), 2010 IEEE International Conference*, pp. 4250–4253, 2010.

[24] M. Kenji, H. Noriyo, K. Noriko, and H. Hajime, "Enhancement of esophageal speech using formant synthesis," *Acoustic. Sci. and Tech.*, pp. 69–76, 2002.

[25] M. P.-M. H. Razo-Chavez, A.; Nakano-Miyatake, "An alaryngeal speech enhancement method based on adpcm approach," *Circuits and Systems, MWSCAS'09, IEEE, International Midwest Symposium*, pp. 1097–1101, Auguest 2009.

[26] B. Garcia, J. Vicente, A. Alonso, and E. Loyo, "Esophageal voices: glottal flow restoration," *Acoustics, Speech and Signal Processing 2005(ICASSP 05)*, pp. 141–144, 2005.

[27] A. Isasi, B. Garcia, and A. M. Zorrilla, "Corrective algorithm for esophageal voice cycle detection," *IEEE*, pp. 150–155, 2011.

[28] M. O. John and B. Garcia, "Quantifiying paramters of a source filter model for oesophageal speech," *IEEE*, pp. 532–53, 2011.

[29] B. Garcia, I. Ruiz, A. Mendez, and M. Mendezona, "Oesophageal voice acoustic parameterization by means of optimum shimmer calculation," *WSEAS Trasactions on Systems*, pp. 489–499, 2008.

[30] B. Garcia, I. Ruiz, J. Vicente, and A. Alonso, "Formants measurement for esophageal speech using wavelet with band and resoultion adjustment," *IEEE Symposium on Signal Processing and Information Technology*, pp. 320–325, 2006.

[31] B. Garcia, J. Vicente, I. Ruiz, A. Alonso, and E. Loyo, "Improvement of esophageal voice's pitch," *Proc. of the 7th Int. Conference on Digital Audio Effects(DAFx04), Italy*, pp. 307–310, 2004.

[32] B. Garcia, I. Ruiz, A. Mendez, and M. Mendezona, "Objective characterization of oesophageal voice supporting medical diagnosis rehabilitation and monitoring," *Computers in Biology and Medicine, Elsevier*, pp. 97–105, 2009.

[33] B. Garcia and A. Mendez, "Oesophageal speech enhancement using poles stablization and kalman filtering," *ICASSP*, pp. 1597–1600, 2008.

[34] O. Ibon, B. Garcia, and Z. M. Amaia, "New approach for oesophageal speech enhancement," *10th International conference, ISSPA*, vol. 5, pp. 225–228, 2010.

[35] S. Schimmel, "Theory of modulation frequency analysis and modulation filtering with applications to hearing devices," Ph.D. dissertation, University of Washington, 2007.

[36] R. Ishaq, M. Shahid, B. Lovstrom, B. G. Zapirain, and I. Claesson, "Modulation frequency domain adaptive gain equlizer using convex optimization," *6th International Conference on Signal Processing and Communication Systems- 2012*, 2012.

[37] M. Shahid, R. Ishaq, B. Sallberg, N. Grbic, B. Lovstrom, and I. Claesson, "Modulation domain adaptive gain equalizer for speech enhancement," in *Signal and Image Processing Application 2011, by IASTED*, 2011.

[38] S. Schimmel and L. E. Atlas, "Analysis of signal reconstruction after modulation filtering," *Advanced Signal Processing Algorithms, Architectures, and Implementations*, vol. 5910, pp. 163–172, 2005.

[39] C. P. Clark and L. Atlas, "A sum-of-product model for effective coherent modulation filtering," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4485–4488, 2009.

[40] C. P. Clark, "Effective coherent modulation filtering and interpolation of long gaps in acoustic signals," Master's thesis, University of Washington, 2008.

[41] P. Clark and L. E. Atlas, "Time-frequency coherent modulation filtering of non-stationary signals," *IEEE transaction on Signal Processing*, vol. 45, no. 57, pp. 4323–4332, 2009.

[42] P. C. Les Atlas and S. Schimmel, "Modulation toolbox version 2.1 for matlab," http://isdl.ee.washington.edu/projects/modulationtoolbox/, September 2010.

[43] A. Alwan. (2012, Feb.) Voicesauce: A program for voice analysis @ONLINE. [Online]. Available: http://www.ee.ucla.edu/ spapl/voicesauce/

[44] S. Boyd and L. Vandenberghe, *Convex Optimization*, 1st ed. Cambridge, UK: Cambridge University Press, 2004.

[45] G. Sell and M. Slaney, "Solving demodulation as an optimization problem," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 8, pp. 2051 –2066, nov. 2010.

Modulation Frequency Domain Adaptive Gain Equalizer Using Convex Optimization

Rizwan Ishaq*, Muhammad Shahid**, Benny Lövström**, Begoña García Zapirain* and Ingvar Claesson**

*Dept. of Elec. Engineering, University of Deusto, Bilbao, Spain
**School. of Elec. Engineering, Blekinge Institute of Technology, Karlskrona, Sweden

***Abstract*—Adaptive gain equalizer (AGE) is a commonly used single-channel speech enhancement algorithm. AGE and its variants has been widely used for speech enhancement applications. There are two broad categories of these variants. The first deals with its improvement in time-frequency domain with readjustment of the used parameters and the second one deals with performing the main filtering operation in modulation frequency domain. This paper evaluates the working of AGE in modulation frequency domain with the use of a demodulation technique which solves the demodulation process as a convex optimization problem. The performance of the modified AGE is compared with the traditional AGE and another modulation frequency domain AGE based on demodulation using the spectral center-of-gravity. These used performance measures are Signal to Noise Ratio Improvement(SNRI), Spectral Distortion(SD) and Mean Option Score(MOS).**

***Index Terms*—Convex demodulation, Center of Gravity, filter bank, Adaptive Gain Equalizer.**

I. Introduction

Different types of background noise corrupts the otherwise clean speech signals in everyday communication. A phone call can be disturbed by a variety of noises present nearby ranging from computer fan noise to factory noise. There have been a variety of methods for reducing noise from speech signal, e.g., spectral subtraction [1] and optimum Wiener filtering [2]. The commonly used method for reducing noise is spectral subtraction but it has an inherent problem of generating musical noise due to spectral flooring [3]. There have also been some efforts to reduce this musical noise such as [4] but this improvement has the tendency of producing audible-distortion causing listening discomfort even compared to the unprocessed signal [5]. Reducing noise without generating artifacts was proposed in [6] but this method fails to address unvoiced speech.

The Adaptive gain equalizer (AGE) is a time domain speech enhancement algorithm in which the speech signal is amplified based on signal-to-noise (SNR) estimates in subbands. A signal is divided into subbands for calculation of a gain which is independent for each band. The algorithm has shown advantages over contemporary techniques because of its low complexity implementation, no requirement of voice activity detector (VAD) and has no presence of musical noise as a result of controlled gains [7]. Additionally, hardware implementations of AGE [8] indicate its importance in speech processing applications.

As an alternative to time domain processing, an implementation of AGE in the modulation domain was presented in a recent study [9]. This method was mainly inspired by the performance advantages of splitting the signal into its frequency bands. The modulation system assumes a speech signal as composed of a modulator and a carrier. Thus the signal is represented by

(1)

where denotes the low frequency part of the signal, called the modulator, that modulates a high frequency carrier . Studies have shown that the modulators of a speech signal are more important for the intelligibility of the speech signals than their counterpart carriers [10]. Modulation systems are based on sub-band modulators and hence perfectly fit the AGE system which works on the sub-bands of the signal. Besides the fact that the study in [9] has reported improvement in performance measures in speech enhancement in comparison to time-domain AGE, the proposed center of gravity (COG) demodulation does not involve an optimization step, the need of which we state in the following.

In this work, we consider AGE in modulation domain by demodulation process as a convex optimization problem presented in [11]. The reason of adaptation of this technique for AGE in modulation domain is mainly the ambiguity associated with the demodulation process of having unlimited number of possible modulator-carrier pairs. Moreover, proven ability of this method for efficiently demodulating a variety of carriers such as harmonic, stochastic and time-varying ones further justifies its usage.

An account of related work in modulation domain and a brief introduction of AGE is provided in Section II. Section III describes a modulation system, a summary of a demodulation technique called spectral center of gravity that used in AGE implementation given in [9]. Section IV starts with an introduction of solving demodulation as an optimization problem and completes with the description of the proposed model of AGE. A comparison of performance of the proposed model is presented in Section V with its time-domain and modulation domain counterparts. Finally, some conclusive remarks about this work are drawn in Section VI with an outline of possible future works in the area.

II. Background

AGE can attenuate noise in speech signals in real time with low computational complexity [12]. It uses an FIR filter bank to divide a speech signal into subbands where speech in each subband is amplified independently. It was also shown that the system can adopt itself for different types of noise. The proposed AGE method using the mixed analog and digital

978-1-4673-2393-2/12/$31.00 ©2012 IEEE

hybrid approach yield around 13 dB speech enhancement [13]. The AGE was originally intended for the digital domain, but [13] provides an analog implementation which does not use quantization and digitization and is best suited for battery powered applications. A hybrid solution to overcome problems related to a digital and an analog implementation of the AGE is found in [14].

Zadeh [15] introduced the modulation domain as a two dimensional bi-frequency system, where time variation of the ordinary frequency is the second dimension. Since then, there have been reasonably large interest in this field for various tasks related to speech processing. Atlas et al. used the concept of coherent modulation for the target talker enhancement in speech enhancement [16]. They proved that working in modulation domain can increase the speech intelligibility. Coherent modulation using the frequency reassignment has been used for speech enhancement and for demodulation of a signal into modulator and carrier [17]. Speech polluted by wind noise has been enhanced by using coherent modulation comb filtering as reported in [18]. Although the modulation filtering has mostly been used for the purpose of speech enhancement, we find some of its applications in audio compression as well [19]. It was showed that a 32 kb/s/channel outperformed MPEG-1 coded at 56 kb/s/channel (both at 44.1 kHz), using the modulation technique.

III. Modulation Domain and AGE

An acoustic spectrum is transformed by short-time Fourier transform into the modulation domain spectrum at a particular acoustic frequency. It has been observed that speech intelligibility can be altered by operating on modulator part of the signal. Shamma [20] reported that auditory cortex neurons possibly decompose the acoustic contents into spectro-temporal modulation contents. It has been found that if the modulators of the speech signal are replaced by constant amplitude modulators, while carriers are preserved, speech does not remain intelligible anymore. However, when the modulators are preserved but carriers are altered, the speech is intelligible [10]. A modulation frequency system is described by the following steps:

Filter bank to get sub-band signals
Demodulation i.e., decomposition of each sub-band signal into a modulator and a carrier.
Analysis of the modulators of the sub-band signals by discrete Fourier transform of each modulators
Modification of the modulators (e.g. linear filtering)
Re-modulation (recombination of modified modulators with original carriers)
Synthesis of signals

The modulation system's filter bank divides the wide-band signal into K narrow-band sub-bands. The signal is passed through the filter bank set of band-pass filters , which renders the sub-band signals .

(2)

s where is convolution operator. The demodulation process decomposes the sub-band signal into its envelope and carrier. It is efficient to decimate the sub-band signals so that the redundant samples may be removed. Modification of the modulators is done by the modulation filtering , i.e., . A modulation spectrogram and modulation analysis can be done by computing the Fourier transform along the time-axis of the spectrogram (magnitude) or by utilizing the spectrum of the envelop signals, which gives the modulation frequency along the horizontal axis and acoustic frequency along the vertical axis. Re-modulation is the process in which modified modulators are combined with the original carriers, obtained in the process of demodulation, to get the modified sub-band signals . The synthesis process reconstructs the modified signal using the modified sub-band signals , according to the following equation. Interpolation must be performed prior to this stage if decimation was done before.

(3)

Following is a brief description on one of the methods used for coherent carrier detection which is also used in this work, apart from convex optimization demodulation process.

A. Spectral Center of Gravity Carrier Estimation

In the Center-of-Gravity(CoG) approach, instantaneous frequency is defined as instantaneous spectrum average frequency of at time [21]. An instantaneous spectrum with short-time Fourier transform is computed as,

(4)

where is a short spectral-estimation window. The instantaneous frequency of the sub-band signal is estimated as,

(5)

The phase of the carrier is computed as follows

(6)

The carrier is

(7)

and the complex valued modulator is given by

(8)

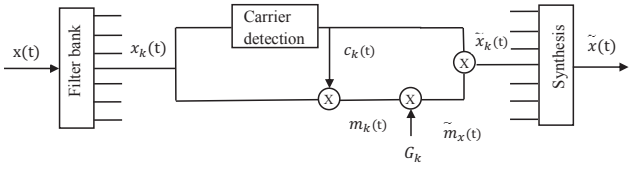

Fig. 1. Adaptive gain equalizer in modulation domain

B. Adaptive Gain Equalizer System

The AGE consists of a filter bank and each sub-band is weighted by a gain function which amplifies the signal when speech is present and keeps the noisy part of the signal, where no speech is present, to unity [7]. A filter bank of K bandpass filters divides the input signal into K sub-bands .

(9)

Here is the impulse response of the filter bank sub-band k and denotes the convolution. The output signal , with the amplified speech signal, is computed as

(10)

where is the AGE weighting function which amplifies the signal when speech is active and is given by

——————— (11)

where is the optimized suppression level for gain function and gain rise exponent constant. is a limiting threshold limiting gain function value. Fast average and slow average of sub-band calculated according to:

(12)

where —— is forgetting factor constant and is sampling frequency.

if

otherwise

(13)

where —— is a positive constant control the noise level. Based on the above mentioned principle of AGE, a speech signal modulator can also be enhanced by the equalizer. Modulation domain separates each sub-band signal into a carrier and a modulator. While only modulators are considered here, the AGE is implemented on each modulator to enhance the speech. The system is shown in figure 1. The mathematics for AGE in the modulation domain is the same as for AGE in the sub-band domain, the long term average and the short term average are calculated for each sub-band modulator, instead of the sub-band itself. The gain function is multiplied with the modulator of the sub-band to yield a modified modulator which is then used with the carrier in the reconstruction stage of the modulation system.

(14)

(15)

The synthesized signal is finally calculated by adding up all the components.

(16)

IV. Convex Optimization and the Proposed Model

One inherent problem with the demodulation technique is the unfortunate presence of unlimited number of possible yet valid modulator-carrier pairs. This predicament can be understood by taking example of a sinusoidal signal that is composed of multiple frequency sinusoids. Such a signal can be decomposed into more than one legitimate modulator and carrier pairs, that are equally correct mathematically. Similar is the case with speech signals when the problem of demodulating it into modulator and carrier is dealt. Thus there is need to add some conditions to the problem which can make the algorithm result into the desired solution. A general optimization problem minimizes a given objective function while fulfilling a set of equality and unequally constraints. If the objective function and inequality constraints are all convex and the equality constraints are all affine, the problem is called a convex optimization problem [22]. Although the modulation problem of equation 1 is not convex as it is, two methods have been suggested in [11] for constraining modulation into convex restrictions. One solution is to work in logarithm domain where the optimization variables can be defined simply as the logarithm of the squared linear optimization variables and . A convex relation is then obtained by just summing the two logarithmic domain variables. The other method of making the problem convex is to work in linear domain where the process involves eliminating the carrier and minimization of only the modulator signal is done. The final expression obtained in linear domain convex optimization is given by the following:

Minimize

where the modulator cost function can be any convex function but the carrier cost function must be both convex and non-decreasing as a requirement of making the problem a convex one. We have followed the linear domain convex optimization method in our work. The interested reader is referred to [11] for detailed analysis of these methods.

V. Comparative Performance Analysis

A. Mean Opinion Score(MOS)

The Mean Opinion Score (MOS) calculated by observing the clean speech signal processed by a system to check how much it degrades the clean speech signal. Fig. 2 shows a female speech signal processed by a system where SNR has been set -10dB for both Engine Noise (EN) and Factory Noise (FN). The system with convex demodulation has MOS value around 3.5 for EN and 3.8 for FN which provides

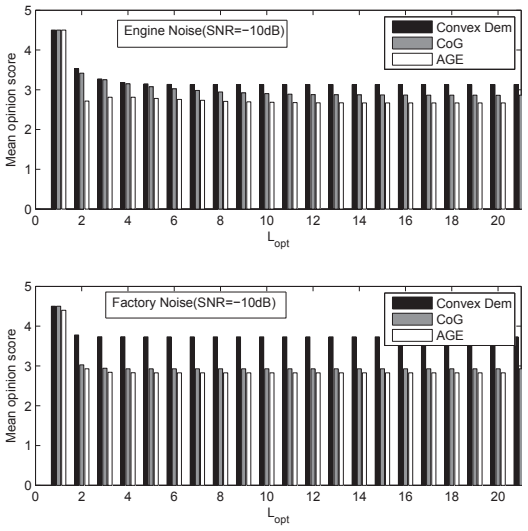

Fig. 2. Mean Opinion Scores(MOS) for all systems with SNR=-10dB

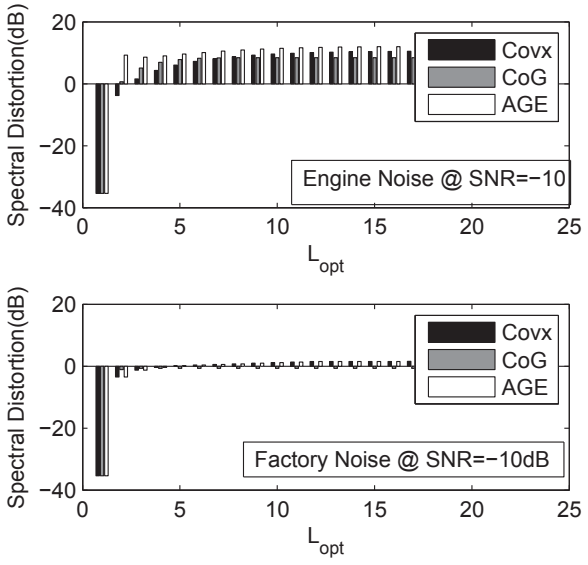

Fig. 3. Spectral Distortion with SNR=-10dB

less degradation as compare to CoG modulation and AGE system where is average MOS observed 3, and less than 3, respectively.

B. Spectral Distortion

Fig.3 shows the Spectral Distortion(SD) for female speech signal contaminated by EN and FN at the SNR of -10dB. The increasing value of increases SD up to 10dB for EN when the system uses AGE while for convex demodulation average SD around 7dB and for CoG demodulation its around 9dB, but for FN, SD for all the system observed around 3dB average.

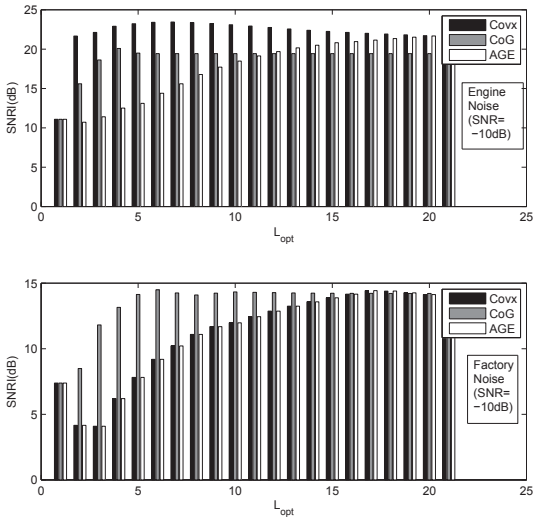

Fig. 4. Signal to Noise Ratio Improvement(SNRI) with SNR=-10dB

C. Signal to Noise Ratio Improvement(SNRI)

Fig. 4 shows the Signal to Noise Ratio Improvement (SNRI) for AGE, MAGE (CoG and Convex demodulation) for a female speech signal distorted by EN and FN having SNR of -10dB. The MAGE methods with convex demodulation has the highest SNRI for all the values of and around 5dB and 8dB improvement over the AGE and MAGE (CoG) methods for EN. But for FN system show improvement after . The MAGE (CoG) in start improved significantly but with increasing value of MAGE (Convex demodulation) has better improvement.

D. Spectrogram Analysis

Fig. 5 and 6 shows spectrogram of original signal with processed signal with AGE, MAGE (convex and CoG demodulation) for FN and EN respectively. The MAGE (convex demodulation) improvement can be observed in term of speech formants being not effected, as visible in spectrogram for both EN and FN.

VI. Conclusion

An alternative method of demodulation has been proposed for AGE in the modulation frequency domain. The presented method solves the demodulation process as a convex optimization problem, thereby avoiding the inherent problem of multiple solutions of a demodulation algorithm. We have tested the proposed method for various conditions and magnitudes of noise injected in a clean speech signal. The performance of our method has been validated by mean opinion score, spectral distortion, signal to noise ratio improvement and spectrogram analysis in comparison to two other techniques.

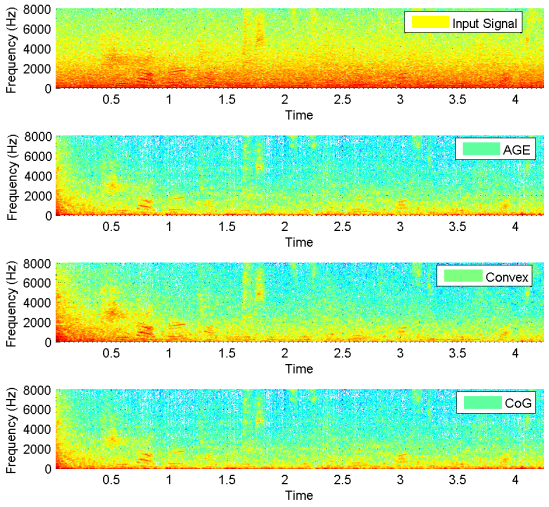

Fig. 5. Spectrogram with Factory Noise(FN) (SNR=-10dB)

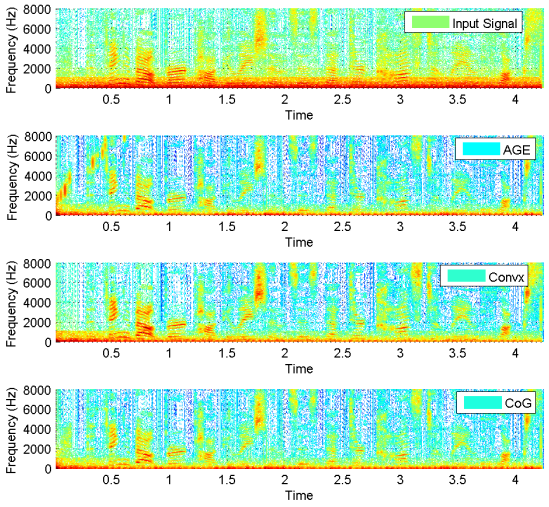

Fig. 6. Spectrogram with Engine Noise(FN) (SNR=-10dB)

REFERENCES

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE trans. Accoust. Speech and Sig. Proc.*, vol. 27, no. 2, pp. 113–120, 1979.

[2] M. H. Hayes, *Statistical Digital Signal Processing and Modeling*, 1st ed. New York, NY, USA: John Wiley & Sons, Inc., 1996.

[3] Z. Goh, K.-C. Tan, and T. Tan, "Postprocessing method for suppressing musical noise generated by spectral subtraction," *Speech and Audio Processing, IEEE Transactions on*, vol. 6, no. 3, pp. 287 –292, may 1998.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, no. 6, pp. 1109 – 1121, dec 1984.

[5] Y. Uemura, Y. Takahashi, H. Saruwatari, K. Shikano, and K. Kondo, "Musical noise generation analysis for noise reduction methods based on spectral subtraction and mmse stsa estimation," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, april 2009, pp. 4433 –4436.

[6] C. Plapous, C. Marro, and P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 6, pp. 2098 –2108, nov. 2006.

[7] N. Westerlund, M. Dahl, and I. Claesson, "Speech enhancement for personal communication using an adaptive gain equalizer," *Elsevier Signal Processing.*, vol. 85, pp. 1089–1101, 2005.

[8] B. Sällberg, N. Grbic, and I. Claesson, "Implementation aspects of the adaptive gain equalizer," 2006.

[9] M. Shahid, R. Ishaq, B. Sällberg, N. Grbic, B. Lövström, and I. Claesson, "Modulation domain adaptive gain equalizer for speech enhancement," in *Signal and Image Processing Application 2011, by IASTED*, 2011.

[10] S. Schimmel, "Theory of modulation frequency analysis with applications to hearing devices," *Ph.D. dissertation*, 2007.

[11] G. Sell and M. Slaney, "Solving demodulation as an optimization problem," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 8, pp. 2051 –2066, nov. 2010.

[12] N. Westerlund, M. Dahl, and I. Claesson, "Real-time implementation of an adaptive gain equalizer for speech enhancement purposes," *WSEAS.*, 2003.

[13] M. Dahl, I. Claesson, B. Sällberg, and H. Akesson, "A mixed analog -digital hybrid for speech enhancement purposes," *ISCAS.*, 2005.

[14] M. Dahl and B. Sällberg, "Speech enhancement implementations in the digital, analog and hybrid domain," *Swedish System on Chip Conference*, 2005.

[15] L. Zadeh, "Frequency analysis of variable networks," in *Proc. IRE*, vol. 38, no. 3, Mar. 1950, pp. 291–299.

[16] S. Schimmel and L. Atlas, "Target talker enhancement in hearing devices," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 31 2008-april 4 2008, pp. 4201 –4204.

[17] S. M. Schimmel, K. R. Fitz, and L. Atlas, "Frequency reassignment for coherent modulation filtering," *IEEE, Acoustics, Speech and Signal Processing, ICASSP*, vol. 5, pp. 261–264, 2006.

[18] B. King and L. Atlas, "Coherent modulation comb filtering for enhancing speech in wind noise," *International Workshop on Acoustice Echo and Noise Control*, Sep 2008.

[19] M. S. Vinton and L. Atlas, "A scalable and progressive audio codec," *IEEE, Acoustics, Speech and Signal Processing, ICASSP*, vol. 5, pp. 3277–3280, 2001.

[20] S. Shamma, "Encoding sound timbre in the auditory system," *IETE J. Res.*, vol. 49, no. 2, pp. 193–205, 2003.

[21] P. Clark and L. E. Atlas, "Time-frequency coherent modulation filtering of non-stationary signals," *IEEE transaction on Signal Processing*, vol. 45, no. 57, pp. 4323–4332, 2009.

[22] S. Boyd and L. Vandenberghe, *Convex Optimization*, 1st ed. Cambridge, UK: Cambridge University Press, 2004.

Adaptive Gain Equalizer for Improvement of Esophageal Speech

Rizwan Ishaq and Begoña García Zapirain
DeustoTech-LIFE, University of Deusto, Spain
Email: rizwanishaq@deusto.es and mbgarciazapi@deusto.es

***Abstract*—The vocal fold modulate the air source from lungs to produce voicing source for speech production. The vocal fold essential part of speech production resides in larynx. The larynx cancer treatment necessitate removal of larynx, in consequences normal speech production destroyed due to no voicing source available. The new voicing source provided artificially or by use of Paryngo-esophageal(PE) segments. The voicing source or residual signal for Esophageal(E) speech uses PE segment, has irregular behavior, and produces degraded quality speech. This paper discussed and evaluated the residual signal or voicing source enhancement of E speech by incorporating speech enhancement method Adaptive Gain Equalizer(AGE), in time-frequency and modulation frequency along with formants modification by Line Spectral Frequencies(LSF) and Linear Predicative Coding(LPC). The system validated by measuring Harmonic to Noise Ratio(HNR) temporally and maximum of 4dB enhancement has been observed in comparison to Kalman filtering based enhancement where enhancement observed maximum of 2dB.**

***Index Terms*—Filter bank, formant enhancement, Esophageal speech, Linear predictive coding, adaptive gain equalizer**

I. Introduction

The normal speech production involves voicing source for speech production and is an essential part. The air source from lungs modulated by vocal folds resides in larynx, for voicing source. But sometime excessive smoking and alcohol use cause laryngeal cancer. The treatment for advanced stage laryngeal cancer is removal of larynx, in consequence, voicing source lost. The other voicing source is needed for speech production, either Pharyngo-esophageal(PE) segment in esophagus or artificial voicing source used. PE is used for both Esophageal (E) speech and Treacheoesophageal(TE) speech, as voicing source but the air source for both methods differ [1]. Electro-larynx(EL) uses external devices for voicing source production which lacks air source. Air source for TE comes from lungs by Tracheoesophagal Puncture(TEP)(a hole between esophagus and trachea), while for E speech air delivered to esophagus through mouth and then released in controlled manner which vibrate PE segment and provides voicing source. But the problem with speech produce either by TE or E speech production methods has low fundamental frequency as well high perturbation in fundamental frequency. The jitter(frequency perturbation), shimmer(amplitude perturbation) are high and Harmonic to Noise Ratio(HNR) low as compared to normal speech due to irregular vibration of PE. Formant frequencies and spectral slope also has different behavior than normal speech [2]. Despite low quality and low intelligible speech due to irregular vibration of PE and low pressure air source for voice source, E speech still preferred methods over other methods because of it does not need surgery and external devices for speech production. The improved quality E speech can be obtained by using signal processing methods to voicing source particularly, by utilizing source filter theory of speech production by decomposing the E speech in to source and filter part by Linear Predicative Coefficients(LPC) analysis.

II. Related Work

In literature, LPC analysis synthesis used to determine vocal tract transfer function and excitation source of E speech and excitation source replaced by LF model for speech quality enhancement [1], [3]. Improvement has been measured of E speech by using LPC and LF model with modified fundamental frequency which has unstable behavior in E speech [4]. The enhanced version of E speech acquired by applying statistical approach to E speech by voice conversion from ES to normal speech [5]. The formant modification based synthesis used for E speech enhancement [6]. The use of Kalman filters along with pole stabilization has been used for enhancement of E speech significantly [7]–[13]. Formant structure modification and excitation source synthesis is used for better quality E speech [14]–[16]. The Adaptive Gain Equalizer(AGE), robust and simple speech enhancement method used for normal speech signal enhancement with lots of variation [17]–[19]. The new signal analysis technique modulation frequency used as well for speech enhancement, speech separation, and recognition [20]–[22].

This paper introduces the system for improving the quality of E speech from database of Spanish vowels \a\, \e\, \i\, \o\, \u\, \bodega\, recorded from speech rehabilitation center from 6 best E speaker with sampling frequency of 44100 Hz. LPC based analysis/synthesis used for residual signal and vocal tract transfer function. AGE in time-frequency and in modulation-frequency domain is used for enhancing residual signal. Formant enhancement is acquired with Line Spectral Frequency(LSF), and new LPC based formant enhancement [23]. The system utilization is measured, temporally by Harmonic to Noise Ratio(HNR), spectrally by spectrogram where formants enhancement observed. The system also provided comparison of HNR with Kalman filtering based E speech enhancement.

978-1-4673-5604-6/12/$31.00 ©2012 IEEE

III. Methodology

LPC analysis is used for decomposing the signal into source and filter part. The speech signal is Pre-Emphasized(pre-emp) for boosting high frequencies and LPC analysis provides us with the residual signal $e(n)$ for speech signal and filter coefficients $a(n)$. Residual signal $e(n)$ passed through AGE and Modulation AGE(MAGE) for reducing noise from $e(n)$, while $a(n)$ passed through formant enhancement methods based on [23]. LPC synthesis used the enhanced residual signal $\hat{e}(n)$ and enhanced filter coefficients $\hat{a}(n)$ to obtained enhanced version of speech signal $\hat{s}(n)$, passed by de-emphasized(de-emp) filter, as shown in Fig.1. The AGE/MAGE and formant modification,

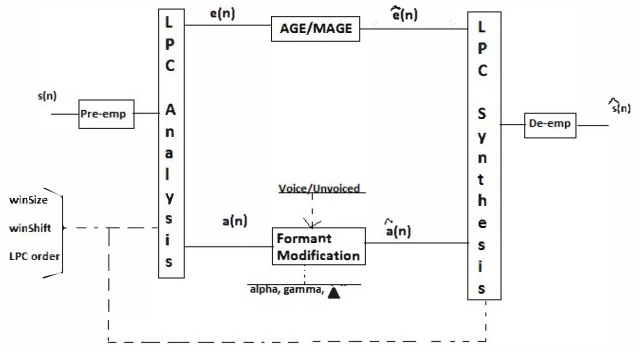

Fig. 1. Simulation setup

different components of system introduces in next sections followed by results and conclusion.

A. Adaptive Gain Equalizer(AGE)

This section introduces the concept of AGE along with the modified version MAGE incorporating the modulation domain [22] [17][1]. Ideally AGE is considered to be robust and low complex speech enhancement methods when compared to other enhancement methods. The AGE method enhanced speech signal quality by raising the SNR of sub-bands of speech signal obtained by applying filter bank. The LPC analysis provide us with residual signals and vocal tract transfer function. The residual signal $e(t)$ obtained from LPC analysis, passed through a filterbank of K bandpass filters, for K sub-bands each denoted by $e_k(n)$.

$$e_k(n) = h_k(n) * e(n) \tag{1}$$

Here $h_k(n)$ is the sub-band k impulse response, $*$ denotes the convolution. The uniformly-spaced sub-bands in a modified Short-Time Fourier Transform(STFT) filterbank is used, where frequency response of each sub-band is roughly-rectangular [22]. The residual signal $e(n)$ modeled as a sum of sub-band signals according to

$$e(n) = \sum_{k=1}^{K} e_k(n) \tag{2}$$

[1]This section based on [17]–[19], [24]

The output modified and enhanced version of the system given by following equation,

$$\hat{e}(n) = \sum_{k=0}^{K-1} g_k(n) e_k(n) \tag{3}$$

where $\hat{e}(n)$ is enhanced voicing source for ES, and $g_k(n)$ is the weighting function for improving signal quality.

1) Gain function: The gain function $g_k(n)$ calculated as the ratio of short term average(fast) $A_k(n)$ and long term average(slow) $B_k(n)$ of k sub-band.

$$g_k(n) = min\left\{\left(\frac{A_k(n)}{L_{opt}.B_k(n)}\right)^{p_k}, L_k\right\} \tag{4}$$

where L_{opt} is the optimized suppression level for gain function, p_k gain rise exponent constant and L_k limiting threshold for gain function. Fast average $A_k(n)$ and slow average $B_k(n)$ of sub-band k calculated according to:

$$A_k(n) = \alpha_k A_k(n-1) + (1-\alpha_k)|e_k(n)| \tag{5}$$

where $\alpha_k = \frac{1}{f_s T_a}$ is forgetting factor constant, f_s and T_a are sampling frequency and time constant respectively.

$$B_k(n) = \begin{cases} A_k(n) & \text{if } A_k(n) \leq B_k(n-1) \\ (1+\beta_k)(B_k(n-1) & \text{otherwise} \end{cases} \tag{6}$$

where $\beta_k = \frac{1}{f_s T_b}$ is a positive constant control the noise level and T_b is a time constant.

B. Modulation Adaptive Gain Equalizer(MAGE)

The modulation frequency domain divides the sub-band signals into modulators and carriers. The modulators are low frequency components of signal while carriers are consider high frequency components. The residual signal sub-bands decomposed into modulators and carriers by demodulation processing by utilization the method based on Center of Gravity(CoG) decomposition [22], [24]. The modulators $m_k(n)$ of residual sub-bands modified according to

$$\hat{e}_k(n) = c_k(n).\hat{m}_k(n) \tag{7}$$

$$\hat{m}_k(n) = m_k(n).g_k(n) \tag{8}$$

The enhanced version $\hat{e}(n)$ signal obtained by synthesis equation (3) and gain for system given by (4) and variables for gain are,

$$A_k(n) = \alpha_k A_k(n-1) + (1-\alpha_k)|m_k(n)| \tag{9}$$

and $B_k(n)$ calculated according to equation (6).

The Fig. 2, and Fig. 3 shows the AGE and MAGE system respectively.

C. Formant Enhancement

The enhancing of formants peaks and spectral valleys significantly improves the quality of speech signal. The formants enhancement methods used in this paper are Line Spectral Frequency(LSF) and new LPC-based formant enhancement taken from [23].

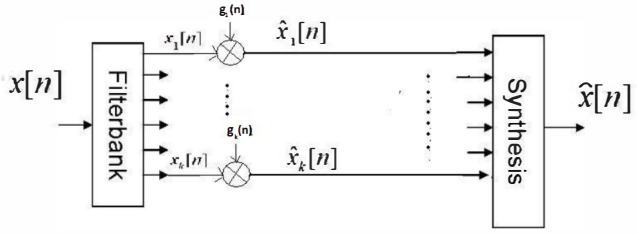

Fig. 2. Adaptive Gain Equalizer

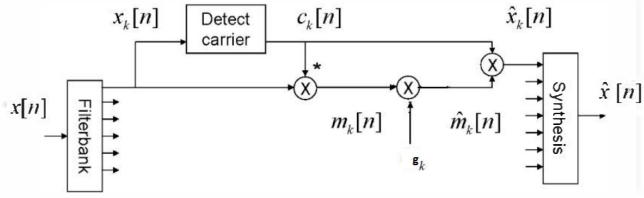

Fig. 3. Modulation Adaptive Gain Equalizer reproduced from [25] used with permission

1) LSF-based Formant Enhancement: LSF based formant enhancement modify LSF positions by shifting LSFs closer to each other for spectral sharpening [23], [26]. The modification of LSFs obtained according to:

$$\hat{lsf}_i = lsf_{i-1}+d_{i-1}+\frac{d_{i-1}^2}{d_{i-1}^2+d_i^2}((lsf_{i+1}-lsf_{i-1})-(d_i+d_{i-1})) \quad (10)$$

here lsf_i and $\hat{lsf}_i$ are original and modified LSF of a frame respectively.

$$d_i = \alpha(lsf_{i+1} - lsf_i) \quad (11)$$

here α is enhancement controlling constant and should be between 0 and 1.

2) LPC-based Formant Enhancement: LPC based formant enhancement modified power spectrum of LPC model and modified power spectrum used to re-evaluating LPC model. Following steps used to modified formants of LPC model [23],

- Power spectrum of LPC
- Low energy part of power spectrum decreased by multiplying small constant γ, while no modification done to formants.
- Re-evaluation of new LPC from the modified power spectrum

LPC based formant enhancement give better results as compare to LSF based formant enhancement.

IV. Simulation Setup

The system tested on the E speech signals recorded from speech rehabilitation center with sampling frequency of 44100 Hz with 6 different people who has good quality E speech. The following E speech Spanish vowel \a\, \e\, \i\, \o\, \u\, \bodega\is used for testing the system. The recorded speech signal down-sampled to 16000 Hz for computation efficiency. The pre-emp filter used with $\alpha = 0.98$. For the LPC analysis, window size for frames set 30ms, with frame overlap of 15ms and order of LPC set 16. The AGE with and without modulation used values shown in table I. The filter bank used 64 number of bandpass filter with decimation factor of 4, although number of filter doesn't effect system behavior. The modification of formants used values for new LPC based enhancement are $\gamma = 0.2$ and $\delta = 200Hz$, while for LSF based enhancement $\alpha = 0.4$ used.

TABLE I
Parameter values for system evolution

Parameter	Value
T_a	30 msec
T_b	3 msec
L_{opt}	$0 \rightarrow 20$
L_k	30 dB
p_k	1

V. Results

A. Spectrogram

The noise between the utterance of vowels has been significantly removed and enhancement of vowel by systems can be observed in the Fig. 6 and Fig. 7 for both MAGE and AGE with $L_{opt} = 6$ for \bodega\. The Fig. 4 shows unprocessed signal spectrogram where noise can be observed between the utterance of words. Fig. 5 shows the spectrogram of processed signal through AGE without decomposing it into source and filter part, although noise reduction has been achieved but formants still mixed and quality of signal still not good. Fig. 6 and 7 shows spectrogram of processed signal where residual signal and formants enhancement applied for MAGE and AGE systems, and improvement in noise reduction as well quality of signal has significant improvement by listening speech signal.

B. Harmonic to Noise Ratio(HNR)

Fig.8 and 9 shows mean Harmonic to Noise Ratio(HNR) in dB for sustained vowels \a\,\e\, \i\o\, \u\and \bodega\. The HNR ratio calculated by using the VoiceSauce [27], by setting the frame size of 30ms and frame step of 15ms and fundamental frequency measurement bounded between 60Hz to 120 Hz because of E speech fundamental frequency fall in this range [2]. The mean values of HNR taken for all frames and for different values of L_{opt}. The results shows improvement after $L_{opt} = 10$, and maximum HNR improvement of 5 dB has been obtained for vowel \o\.

C. Comparision with Kalman Filtering approach

The Kalman filtering and poles stabilization is used for enhancing E speech by applying Kalman filtering to E speech without decomposing into source and filter and then modified

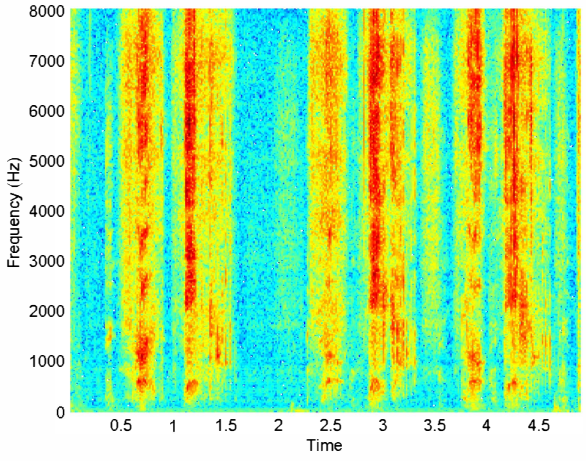

Fig. 4. Unprocessed E speech signal(bodega)

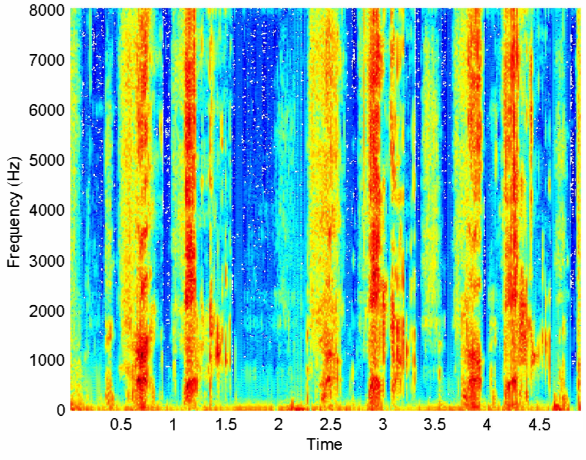

Fig. 5. Processed E speech signal (without decomposing into source and filter part),$L_{opt} = 6$

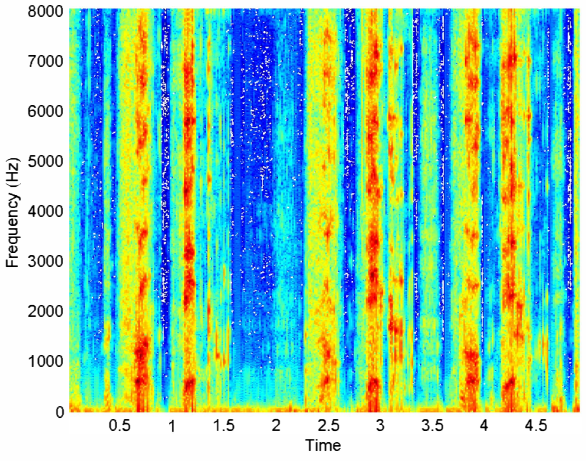

Fig. 6. Processed E speech signal (Modulated AGE),$L_{opt} = 6$

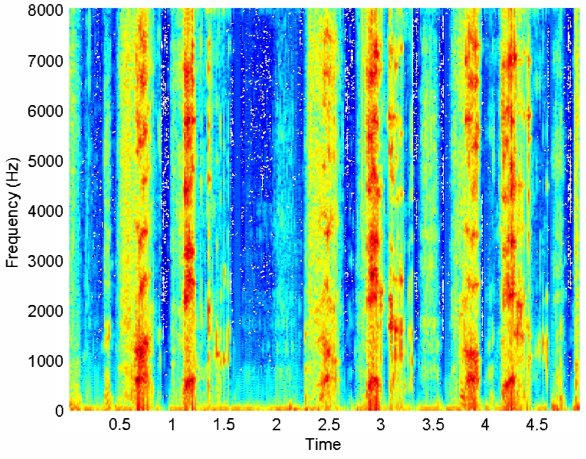

Fig. 7. Processed E speech Signal (AGE),$L_{opt} = 6$

poles by stabilization it as given in [7]. The average enhancement in HNR has been observed around 2dB for most of cases, in comparison MAGE and AGE produces enhancement around 4dB which can be observed in Figs.8,9.

VI. Conclusion

The improved HNR has validated the system capability to improve quality of E speech. The system successfully removed noise, as well tried to enhanced residual signal for better and intelligible E speech signal. The MAGE modulation frequency system provides better enhancement in comparison to AGE and overall both systems outperformed Kalman filtering based system [7]. Along with voicing source enhancement, formant enhancement significantly improves quality of E speech in comparison to poles modification by shifting upward [7] when conducted listening test.

Although LPC analysis/synthesis provides good estimation of voicing source signal, but as comparison to normal speech this estimation in not perfect and modeling of PE segment can provides better voicing source signal which can be improved and modified through this system. The future can be to provide optimized value of L_{opt} by having noise information from modeling of PE segment.

Acknowledgment

This research was partially granted by Deiker of Deusto University and Department of Education and Researcher of Basque government under the support of Duestotech eVida group of Deusto university project.

References

[1] Q. Yingyong, "Replacing tracheoesophageal voicing source and lpc synthesis," *Acoustical Society of America*, vol. 5, pp. 2461–2465, 1995.
[2] E. Lundstrom, "Voice function and quality of life in laryngectomees," Ph.D Thesis, Karolinska Institutet, Stockholm, Sweden, 2009.

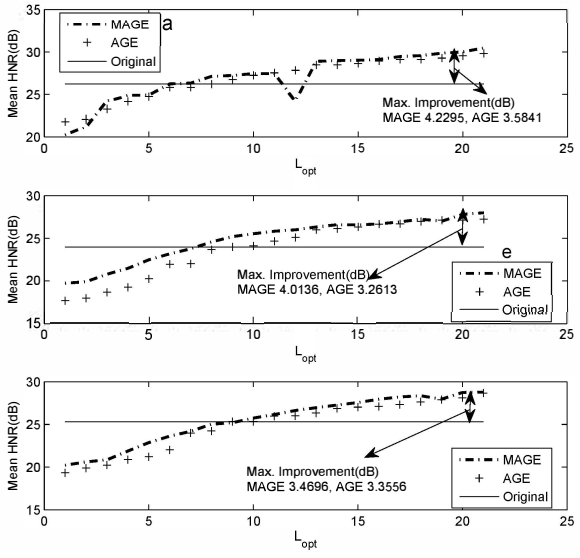

Fig. 8. HNR for \a\, \e\,\i\

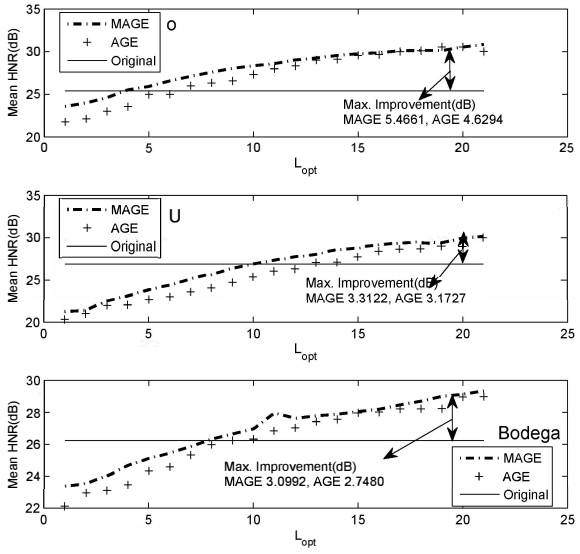

Fig. 9. HNR for \o\, \u\,\bodega\

[3] Q. Yingyong, W. Bernd, and B. Ning, "Enhancement of female esophageal and tracheoesophageal speech," *Acoustical Society of America*, vol. 3, pp. 1228–1235, 1990.

[4] R. Sirichokswad, P. Boonpramuk, N. Kasemkosin, P. Chanyagorn, W. Charoensuk, and H. H. Szu, "Improvement of esophageal speech using lpc and lf model," *Internation Conf. on Biomedical and Pharamaceutical Engineering 2006*, pp. 405–408, 2006.

[5] H. Doi, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Statistical approach to enhancing esophageal speech based on gaussian mixture models," *Acoustics Speech and Signal Processing(ICASSP)*, pp. 4250–4253, 2010.

[6] M. Kenji and H. Noriyo, "Enhancement of esophageal speech using formant synthesis," *Acoustics, Speech and Signal Processing, International conf.*, pp. 81–85, 1999.

[7] B. Garcia and A. Mendez, "Oesophageal speech enhancement using poles stablization and kalman filtering," *ICASSP*, pp. 1597–1600, 2008.

[8] O. Ibon, B. Garcia, and Z. M. Amaia, "New approach for oesophageal speech enhancement," *10th International conference, ISSPA*, vol. 5, pp. 225–228, 2010.

[9] B. Garcia and J. Vincente, "Software for measuring and improving esophageal voice," *Internation Conference on Digital Audio Effects(DAFx04), Italy*, vol. 5, pp. 303–306, 2004.

[10] B. Garcia, I. Ruiz, A. Mendez, and M. Mendezona, "Oesophageal voice acoustic parameterization by means of optimum shimmer calculation," *WSEAS Trasactions on Systems*, pp. 489–499, 2008.

[11] B. Garcia, I. Ruiz, J. Vicente, and A. Alonso, "Formants measurement for esophageal speech using wavelet with band and resoultion adjustment," *IEEE Symposium on Signal Processing and Information Technology*, pp. 320–325, 2006.

[12] B. Garcia, I. Ruiz, A. Mendez, and M. Mendezona, "Objective characterization of oesophageal voice supporting medical diagnosis rehabilitation and monitoring," *Computers in Biology and Medicine, Elsevier*, pp. 97–105, 2009.

[13] B. Garcia, J. Vicente, A. Alonso, and E. Loyo, "Esophageal voices: glottal flow restoration," *Acoustics, Speech and Signal Processing 2005(ICASSP 05)*, pp. 141–144, 2005.

[14] R. H. Ali and S. B. Jebara, "Esophageal speech enhancement using excitation source synthesis and formant structure modification," *IEEE*, 2005.

[15] R. A. Prosek and L. L. Vreeland, "The intelligibility of time domain edited esophageal speech," *American Speech Language Hearing Association*, vol. 44, pp. 525–534, 2001.

[16] A. Loscos and J. Bonada, "Esophageal voice enhancement by modeling radiated pulses in frequency domain," *Audio Engineering Society*, 2006.

[17] S. Muhammad, I. Rizwan, S. Benny, G. Nedelko, L. Benny, and C. Ingvar, "Modulation domain adaptive gain equalizer for speech enhancement," *IASTED International Conference on Signal and Image Processing and Applications*, 2011.

[18] N. Westerlund, M. Dahl, and I. Claesson, "Speech enhancement for personal communication using an adaptive gain equalizer," *EURASIP Journal on Audio Speech and Music Processing*, pp. 1089–1101, 2005.

[19] B. et al, "An improved adaptive gain equalizer for noise reduction with low speech distortion," *EURASIP Journal on Audio Speech and Music Processing*, 2011.

[20] P. Clark and L. Atlas, "Time-frequency coherent modulation filtering of nonstationary signals," *IEEE Trans. Signal Processing*, vol. 57, pp. 4323 – 4332, 2009.

[21] G. Sell and M. Slaney, "Solving demodulation as an optimization problem," *IEEE Transcations on Audio, Speech and Language Processing*, 2010.

[22] S. M. Schimmel, "Theory of modulation frequency analysis and modulation with application to hearing devices," Ph.D Thesis, University of Washington, 2007.

[23] T. Raitio, H. P. A. Suni, M. Vainio, and P. Alku, "Comparison of formant enhancement methods for hmm-based speech synthesis," *7th ISCA Speech Synthesis workshop, SSW7*.

[24] R. Ishaq, "Adaptive Gain Equalizer and Modulation Frequency Domain for Noise Reduction," Master Thesis, Blekinge Institute of Technology, Karlskrona, Sweden.

[25] C. P. Clark, "Effective coherent modulation filtering and interpolation of long gaps in acoustic signals," Master Thesis, University of Washington, USA, 2008.

[26] Z. Ling, Y. Wu, L. Qin, and R. Wang, "Ustc system for blizzard challenge 2006 an improved hmm-based speech synthesis method," *Blizzard Challenge Workshop*, 2006.

[27] A. Alwan. (2012, Feb.) Voicesauce: A program for voice analysis @ONLINE. [Online]. Available: http://www.ee.ucla.edu/ spapl/voicesauce/

MODULATION DOMAIN ADAPTIVE GAIN EQUALIZER FOR SPEECH ENHANCEMENT

Muhammad Shahid, Rizwan Ishaq, Benny Sällberg, Nedelko Grbic, Benny Lövström and Ingvar Claesson
Department of Signal Processing, Blekinge Institute of Technology, SE-37179 Karlskrona, Sweden
Correspondence author: muhammad.shahid@bth.se

ABSTRACT
This paper evaluates speech enhancement by filtering in the modulation frequency domain, as an alternative to filtering in conventional frequency domain. Adaptive Gain Equalizer (AGE) is a commonly used single-channel speech enhancement algorithm. A recently introduced class of signal transformations called modulation transform has successfully made its place alongside classical time/frequency representations. This paper presents an implementation of AGE within modulation system, for the purpose of enhancing the speech signal. The successful implementation of the proposed system has been validated with various performance measurements, i.e., Signal to Noise Ratio Improvement (SNRI), Mean Opinion Score (MOS) and Spectral Distortion (SD). A spectrogram analysis is also presented to further substantiate the performance of this work.

KEY WORDS
Speech enhancement, Adaptive gain equalizer, Modulation domain.

1 Introduction

Speech as the main part of the communication systems, is usually degraded during the transmission by different types of noise, e.g., Gaussian noise, engine noise, periodic noise and other interferences. There are a variety of methods for reduction of noise from speech signal, e.g., spectral subtraction (frequently used for noise reduction) [1] and optimum Wiener filtering [2]. Adaptive Gain Equalizer (AGE) [3] is a noise reduction method that focuses on enhancing the speech signal instead of suppressing the noise. The speech enhancement is carried out by weighting the sub-bands in time-frequency domain according to an estimate of the Signal-to-Noise Ratio (SNR). This method offers better result in terms of low complexity, low delay, low distortion and there is no need for Voice Activity Detector (VAD) .
The modulation system assumes that a speech signal is composed of a modulator and a carrier. The signal is represented by,

$$x(t) = m(t)c(t) \tag{1}$$

where $m(t)$ denotes the low frequency part of the signal, called modulator, and it modulates a high frequency carrier $c(t)$. Studies have shown that the modulators of speech signal are most important for the intelligibility of the speech signal. The importance of the modulator in speech signals brought the attention of many researchers .
AGE implementation has been intended so far in time-frequency domain, but here an implementation of AGE in a modulation system is proposed. Modulation systems which are based on sub-band modulators, perfectly fit the AGE system which works on the sub-bands of the signal.

1.1 Literature Survey

Zadeh [4] is considered to be the pioneer of the field of modulation domain who suggested a two dimensional bi-frequency system, where time variation of the acoustic frequency is the second dimension of frequency. Atlas et al. used the concept of coherent modulation for the target talker enhancement in speech enhancement [5]. They proved that modulation domain moderately increases the speech intelligibility . Coherent modulation using the frequency reassignment has been used for speech enhancement and for demodulation of a signal into modulator and carrier [6]. Li et al. described the theory behind modulation filtering which offers a new approach to modifying non-stationary signals e.g., speech. They presented the coherent modulation analysis based on instantaneous frequency estimation using conditional mean frequency. In addition, they showed that the proposed method accurately estimates the carriers and modulators of the signals [7]. Speech polluted by wind noise has been enhanced by using coherent modulation comb filtering by King et al. [8]. Although the modulation filtering has mostly been used for the purpose of speech enhancement, Vinton et al. also used it for audio compression. They showed that a 32 kb/s/channel outperformed MPEG-1 coded at 56 kb/s/channel (both at 44.1 kHz), using the modulation technique [9]. The concept of homomorphic demultiplication is connected to the modulation spectral analysis/synthesis and it was outlined by Atlas et al. in [10]. Clark et al. showed in [11] the effectiveness of modulation filtering by measuring the empirical modulation frequency response and got a near-ideal response performance, and 25 dB improvement has been shown for suppressing undesired modulation frequencies over incoherent modulation. Clark presented the Center of Gravity (COG) method for decomposition of a sub-band signal, and he used coherent modulation filtering for the interpolation

of long gaps in acoustic signals [12].
The concept of AGE for the reduction of noise in speech signals, has shown its success in real time and proven to be a low complexity system [3]. The method used an FIR filter bank to get the required results and it was also shown that the system adapted itself for different types of noise. The proposed AGE method using the mixed analog and digital hybrid approach yielded around 13 dB speech enhancement [13]. The AGE was originally intended for the digital domain, but [14] provides an analog implementation which does not use quantization and digitization and it is also best fitted for battery powered applications. A hybrid solution to overcome problems related to a digital and an analog implementation of the AGE is found in [15].

1.2 Main Contribution

The main contribution of this paper is to combine the AGE and modulation system domain for speech enhancement. Hence, the advantage of benefits from both of the fields has been taken to build up a new system. This approach has proven to be robust, flexible in implementation and has been validated by performance measures like Signal to Noise Ratio Improvement (SNRI), Mean Opinion Score (MOS) and Spectral Distortion (SD). Section 2 briefly introduces the modulation system, section 3 introduces the concept of AGE and its operation in the modulation frequency domain and section 4 evaluates the proposed system. Section 5 concludes this work with a summary and future research directions in the area.

2 Modulation System

A modulation domain spectrum is obtained from a certain acoustic spectrum by taking short-time Fourier transform (STFT) of the speech signal at the given acoustic frequency. The speech signal modulators are the most important components for speech intelligibility. Shamma [16] reported that auditory cortex neurons possibly decompose the acoustic contents into spectro-temporal modulation contents. It has been found that if the modulators of the speech signal are replaced by constant amplitude modulators, while carriers are preserved, speech is not intelligible. However when the modulators are preserved but carriers are altered, the speech is intelligible [17]. Modulation domain actually decomposes the speech, or other natural signals, into modulators and carriers whereafter the modulators of the signals are analyzed. A general framework for modulation frequency domain analysis, and filtering is given in figure 1. A modulation frequency system is described by the following steps:

- Filter bank to get sub-band signals
- Demodulation i.e., decomposition of each sub-band signal into a modulator and a carrier.

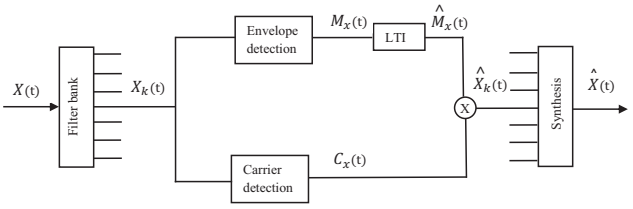

Figure 1. A general framework of the modulation filtering and analysis system [17]

- Analysis of the modulators of the sub-band signals by discrete Fourier transform of each modulators
- Modification of the modulators (e.g. linear filtering)
- Re-modulation (recombination of modified modulators with original carriers)
- Synthesis of signals

The modulation system filter bank divides the wide-band signal into K narrow-band sub-bands. The signal $x(t)$ is passed through the filter bank's set of band-pass filters h_k, which renders the sub-band signals $x_k(t)$.

$$x_k(t) = h_k * x(t) \tag{2}$$

where $*$ denotes the convolution operator. The demodulation process decomposes the sub-band signal into its envelope and carrier. Its efficient to decimate the sub-band signals so that the redundant samples may be removed. Modification of the modulators is done by the modulation filtering which mostly uses linear time invariant filters $g(t)$, i.e., $\hat{m_k}(t) = m_k(t)g(t)$. A modulation spectrogram and modulation analysis can be done by computing the Fourier transform along the time-axis of the spectrogram (magnitude) or by utilizing the spectrum of the envelop signals, which gives the modulation frequency along horizontal axis and acoustic frequency along vertical axis. Re-modulation is the process in which modified modulators $\hat{m_k}(t)$ are combined with the original carriers, obtained in the process of demodulation, to get the modified sub-band signals $\hat{x}_k(t)$. The synthesis process reconstructs the modified signal $\hat{x}(t)$ using the modified sub-band signals $\hat{x}_k(t)$, according to the following equation. Interpolation must be performed prior to this stage if decimation was done before.

$$\hat{x}(t) = \sum_{k=1}^{K} \hat{x}_k(t) \tag{3}$$

Envelope detection is used for demodulation of a signal and it is the most important part of the modulation frequency system. There are two types of envelope detectors mostly used, coherent envelope detection and incoherent envelope detection. Magnitude, or magnitude-like, operations are used to estimate modulators in incoherent detection, while coherent detection use the carrier estimate operations. Incoherent envelope detection detects the envelope

and carrier independently and coherent detection uses the carrier estimation for the calculation of the envelope. Following is a brief description about one of the methods used for coherent carrier detection which is used in this work.

2.1 Spectral Center of Gravity Carrier Estimation

In this recently introduced method of the center-of-gravity approach, instantaneous frequency $\omega_k(n)$ is defined as instantaneous spectrum average frequency of $x_k(t)$ at time t [18]. An instantaneous spectrum with short-time Fourier transform is computed as,

$$S_k(\omega, t) = \sum_p g(p) x_k(t+p) e^{-j\omega p} \tag{4}$$

where $g(p)$ is a short spectral-estimation window. The instantaneous frequency $\omega_k(t)$ of the sub-band signal $x_k(t)$ is estimated as,

$$\omega_k(t) = \frac{\int_{-\pi}^{\pi} \omega |S_k(\omega, t)|^2 d\omega}{\int_{-\pi}^{\pi} |S_k(\omega, t)|^2 d\omega} \tag{5}$$

The phase $\phi_k(t)$ of the carrier is computed as follows

$$\phi_k(t) = \sum_{p=0}^{t} \omega_k(p) \tag{6}$$

The carrier c_k is

$$c_k(t) = e^{j\phi_k(t)} \tag{7}$$

and the complex valued modulator $m_k(t)$ is given by

$$m_k(t) = x_k(t) c_k^*(t) \tag{8}$$

3 Adaptive Gain Equalizer System

As discussed in [3], the AGE consists of a filter bank with different band-pass filters. Each sub-band is weighted by a gain function which amplifies the signal when speech is present and keeps the noisy part of the signal, where no speech is present, to unity. A filter bank of K bandpass filters divides the input signal $x(n)$ into K sub-bands $x_k(n)$.

$$x_k(n) = h_k * x(n) \tag{9}$$

Here h_k is the impulse response of the filter bank sub-band k and $*$ denotes the convolution. The time domain signal is modeled as a sum of sub-band signals, according to:

$$x(n) = \sum_{k=1}^{K} x_k(n) = \sum_{k=1}^{K} (s_k(n) + w_k(n)) \tag{10}$$

where $s_k(n)$ is the desired speech signal related to k^{th} sub-band, while $w_k(n)$ is the additive noise in the sub-band k.

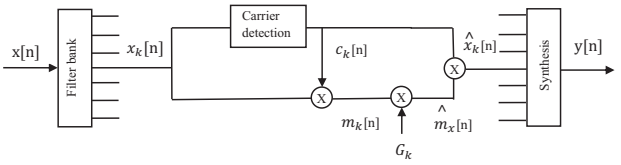

Figure 2. Adaptive gain equalizer in modulation domain

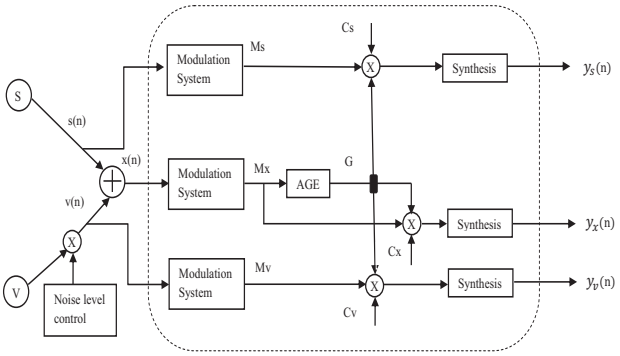

Figure 3. Experiment setup

The output signal $y(t)$, with the amplified speech signal, is computed as

$$y(n) = \sum_{k=1}^{K} G_k(n) x_k(n) \tag{11}$$

where $G_k(n)$ is the AGE weighting function which amplifies the signal when speech is active.

3.1 Gain Function

Two terms used for the calculation of the gain function are; a long term (slow) average $A_{s,k}(t)$ and the short term (fast) average $A_{f,k}(t)$. The short term average, for sub-band k, $A_{f,k}(n)$ is calculated as

$$A_{f,k}(n) = \alpha_k A_{f,k}(n-1) + (1-\alpha_k) \mid x_k(n) \mid \tag{12}$$

where α_k is a small positive constant, given by

$$\alpha_k = \frac{1}{T_{s,k} F_s} \tag{13}$$

where F_s is the sampling frequency in Hz and $T_{s,k}$ is a time constant in seconds. In the same manner, a slow average is computed as

$$A_{s,k}(n) = (1+\beta_k) A_{s,k}(n-1) \tag{14}$$

if $A_{s,k}(n-1) \leq A_{f,k}(n)$, and

$$A_{s,k}(n) = A_{f,k}(n) \tag{15}$$

if $A_{s,k}(n-1) > A_{f,k}(n)$
where β_k is a small positive constant. The AGE gain function is computed as:

$$G_k(n) = \left(\frac{A_{f,k}(n)}{A_{s,k}(n)}\right)^{p_k} \quad (16)$$

where $p_k \geq 0$, and $A_{s,k}(n) > 0$.

3.2 Modulation Domain AGE

The functionality of the AGE has been extended to work in the modulation domain for speech enhancement. Modulation domain separates each sub-band signal into a carrier and a modulator. While only modulators are considered here, the AGE is implemented on each modulator to enhance the speech. The system is shown in figure 2. The mathematics for AGE in the modulation domain is the same as for AGE in the sub-band domain, the long term average and the short term average are calculated for each sub-band modulator, instead of the sub-band itself. The gain function is multiplied with the modulator of the sub-band to yield a modified modulator $\hat{m}_k(n)$ which is then used with the carrier in the reconstruction stage of the modulation system.

$$\hat{m}_k(n) = m_k(n)G_k \quad (17)$$
$$\hat{x}_k(t) = c_k(n)\hat{m}_k(n) \quad (18)$$

The synthesized signal $y(n)$ is finally calculated by adding up all the components.

$$y(n) = \sum_{k=1}^{K} \hat{x_k}(n). \quad (19)$$

The gain function G_k is given by

$$G_k = \min\left(L, \frac{A_{f,k}}{L_{opt}.A_{s,k} + \epsilon}\right) \quad (20)$$

where $A_{f,k}$ denotes short term average and $A_{s,k}$ denotes the long term average, L is a limiting threshold which limits the gain function's value and L_{opt} is an optimum level of control on the value of the gain function. The averages are computed as:

$$A_{f,k}(n) = \alpha_f A_{f,k}(n-1) + (1-\alpha_f) \mid m(n) \mid \quad (21)$$
$$A_{s,k}(n) = \alpha_s A_{s,k}(n-1) + (1-\alpha_s) \mid m(n) \mid \quad (22)$$
$$A_{s,k}(n) = \min\left(A_{s,k}(n), A_{f,k}(n)\right) \quad (23)$$

where α_f and α_s are time constants of the short term and long term averages, respectively.

4 Evaluation of The Proposed System

Figure 3 shows the experimental setup, where $s(n)$ is the clean speech signal, $v(n)$ is a noise signal and $x(n)$ is the sum of speech and noise signals $(s(n) + 10^{\frac{-SNR}{20}} v(n))$

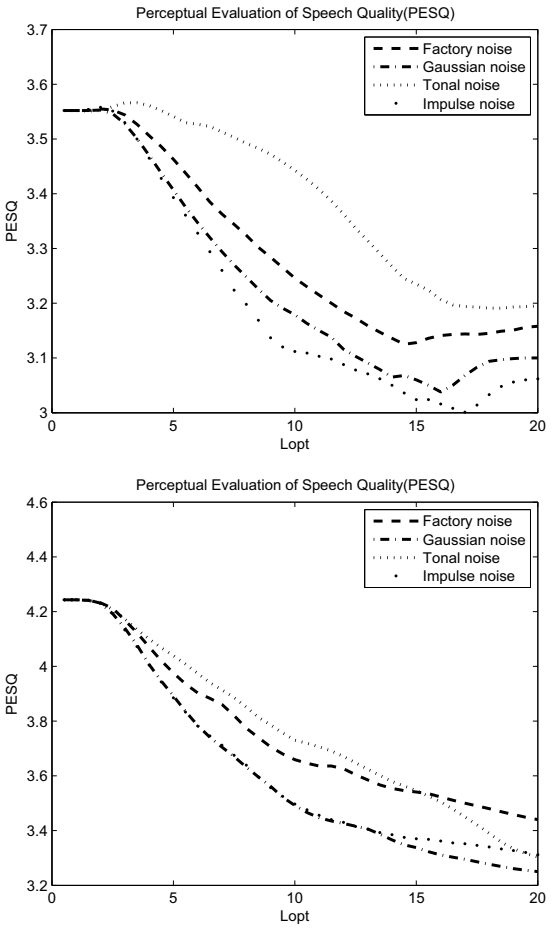

Figure 4. MOS for the processed male speech signal (upper) and female speech signal (lower) with noise at 10 dB SNR

scaled by desired level of Signal to Noise Ratio (SNR). M_s, C_s, M_x, C_x, M_v and C_v are the signal matrices of modulators and carriers for $s(n)$, $x(n)$ and $v(n)$ respectively. The gain matrix G is calculated by passing M_x through AGE system. This G is then multiplied with the M_x, M_s and M_v, whereafter the re-modulation and the synthesis processes generate the output signals $y_x(n)$, $y_s(n)$, $y_v(n)$, as depicted in figure 3. The system was evaluated with the following parameter setings. $L = 1$, L_{opt}= 1 to 20, $T_s = 4s$ and $T_f = 0.04s$. The speech signals comprise male F_s=16 kHz and female F_s=16 kHz speech signals and the noise signals are scaled so as to have 10 dB, 5 dB, 0 dB and -5 dB SNR. Noise signals used were Engine Noise (EN), Factory Noise (FN), Gaussian Noise (GN), Tonal Noise (TN) and Impulse Noise (IN). The performance measurement was evaluated by the Signal to Noise Ratio Improvement (SNRI), Perceptual Evaluation of Speech Quality (PESQ) and Spectral Distortion (SD). SNRI of male speech signal for TN at 0 dB SNR with $L_{opt} = 20$ was around 10 dB and for other noises was between 4 dB to 6

Table 1. Spectral distortion (SD) results

Noise SNR	0 dB		10 dB	
L_{opt} **range**	0 to 5	5 to 20	0 to 5	5 to 20
Speaker	Male			
SD for FN [dB]	-18 to -4	-4 to -2	-18 to -4	-4 to -2
SD for IN [dB]	-18 to -4	-4 to -2	-18 to -4	-4 to -2
SD for TN [dB]	-18 to -6	-4 to -2	-18 to -4	-6 to -2
Speaker	Female			
SD for FN [dB]	-34 to -12	-12 to -2	-34 to -15	-15 to -4
SD for IN [dB]	-34 to -15	-15 to 0	-18 to -6	-6 to -2
SD for TN [dB]	-34 to -15	-15 to -7	-34 to -18	-18 to -8

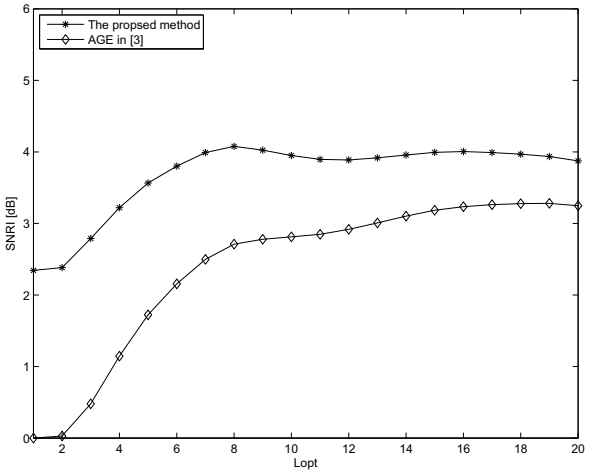

Figure 5. SNRI plots of two speech enhancement methods

dB. The female speech signal also had SNRI of 9 dB for TN and around 3 to 5 dB for EN, FN, GN, IN at 0d B SNR. PESQ has been calculated by comparing $s(n)$ and $y_s(n)$ which gives an objective measure of how much degradation the system has introduced on the speech signal due to introducing the AGE gain function. The objective Mean Opinion Score (MOS) as computed by the PESQ for most of the tests given above was 3, which is considered fair for speech signals. Experiments have been performed to find out the optimal value on the critical system parameter L_{opt}, for different noise cases and for different speaker situations. Figure 4 shows the MOS values for both male and female speech signals at 10 dB of noise SNR. It is interesting to note that female speech has higher values of MOS than male speech under similar conditions. This observation is attributed to the fact that female speech with higher pitch is less affected by some noises. Moreover, the SD is very low for $L_{opt} < 5$ and then increases rapidly with increasing L_{opt} values for all tests. For male speech signal, the SD at L_{opt} = 20 is around -2 dB and -4 dB for FN,GN,TN and IN and some of them are shown in table 1. The female speech signal has different behavior than the male speech signal on SD. For female speech, SD is found to be -2 dB for EN, GN, IN and -4 dB for FN and -8 dB for TN at the L_{opt}=20.

The proposed method was also compared against the speech enhancement method by AGE proposed in [3]. It was observed that the proposed method has better performance than the reference method of [3]. One such comparison is shown in figure 5 where a male speech signal having mixed with 5 dB SNR factory noise is enhanced by two methods and the proposed method clearly outperforms its counterpart in [3].

4.1 Spectrogram Analysis

The spectrogram of a male speech signal that has been mixed with gaussian noise at 10 dB SNR and the spectrogram after enhancement with the propsed AGE system, are given in figure 6. The AGE algorithm converges after 0.2 seconds for all test cases, whereafter it may be observed that the disturbing noise is reduced while the formants of the speech are maintained. Enhanced signal $y_x(n)$ has shown the formants very clearly after the processing. Although the Gaussian noise is spread throughout the frequency plane, the AGE works very efficiently, but a little bit speech signal energy has also been lost. The spectrogram of male speech signal mixed with tonal noise at 0 dB SNR and enhanced male speech signal by the AGE was also observed. The tonal noise which had all of its energy around 1 kHz has been reduced by the AGE, i.e., reduced its energy, while maintaining the formants of speech. Moreover, the impulse noise at 0 dB SNR, which is similar to gaussian noise in spreading its energy through all the frequencies, has been successfully eliminated.

5 Conclusion

A novel approach of speech enhancement in modulation frequency domain has been explored and the promising results obtained by using the proposed method have been presented in this paper. The adaptive gain equalizer (AGE), which has shown its advantages already in digital, analog

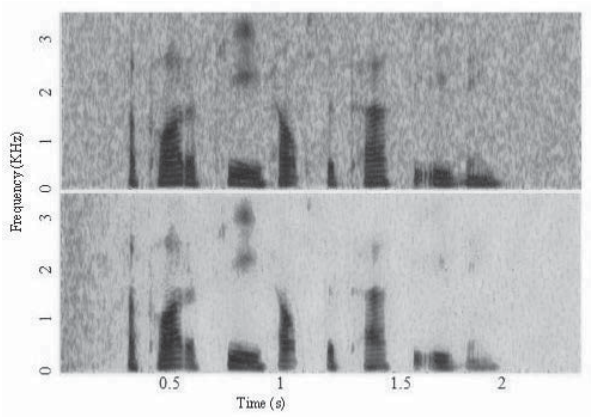

Figure 6. Spectrogram of noisy male speech (upper) having Gaussian noise at 10 dB SNR and the enhanced signal by the proposed method (lower)

and hybrid domains by its simplicity, low complexity for being robust to different noisy environments, has been implemented in the modulation frequency domain in this paper. The detailed analysis of the system has put light on its advantages and disadvantages, i.e. where the evaluation section highlights the compromise between low SD and high SNRI. The system provides good improvement on the female speech signal, with better SNRI, low SD, fair MOS, and output speech signal sounds good. The maximum SNRI obtained for the female speech signal analysis was approximately 9 dB and SD of female speech for some noise has been shown 0 dB.

The spectrogram analysis provides another view of these results. The AGE gain function adapts during the first 0.2 seconds. This start-up time can be reduced by varying the integration time, but changing the integration time has obvious consequences on the signal integrity and the noise reduction performance. Moreover, the proposed method has shown its potential as a better alternative to the traditional methods of speech enhancement.

Future work is to implement this system in real time and other speech enhancement methods may also be tried in modulation domain.

References

[1] S. F. Boll, Suppression of acoustic noise in speech using spectral subtraction, *IEEE trans. Accoust. Speech and Sig. Proc.*, 27(2), 1979, 113-120.

[2] M. H. Hayes, *Statistical Digital Signal Processing and Modeling* (New York: John Wiley and Sons Inc., 1996).

[3] N. Westerlund, M. Dahl and I. Claesson, Real-time implementation of an adaptive gain equalizer for speech enhancement purposes, *Proc. 2nd WSEAS International Conf. on Electronics, Control and Signal Processing*, Singapore, 2003, 2:1–2:8.

[4] L. Zadeh, Frequency analysis of variable networks, *Proc. IRE*, 38(3), 1950, 291-299.

[5] L. E. Atlas and S. M. Schimmel, Target talker enhancement in hearing devices, *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, USA, 2008, 4201-4204.

[6] S. M. Schimmel, K. R. Fitz and L. E. Atlas, Frequency reassignment for coherent modulation filtering , *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Toulouse, France, 2006, 261-264.

[7] Q. Li and L. Atlas, Coherent modulation filtering for speech, *Proc. IEEE, International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, USA, 2008, 4481-4484.

[8] B. King and L. Atlas, Coherent modulation comb filtering for enhancing speech in wind noise, *Proc. International Workshop on Acoustice Echo and Noise Control*, Seattle, USA, 2008.

[9] M. S. Vinton and L.E. Atlas, A scalable and progressive audio codec, *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Salt Lake City, USA, 2001, 3277-3280.

[10] L. Atlas, Q. Li and J. Thompson, Homomorphic modulation spectra, *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Montreal, Canada, 2004, 761-764.

[11] C. P. Clark and L. Atlas, A sum-of-product model for effective coherent modulation filtering, *Proc. IEEE, International Conference on Acoustics, Speech and Signal Processing*, Taipei , China, 2009, 4485-4488.

[12] C. P. Clark, Effective coherent modulation filtering and interpolation of long gaps in acoustic signals, *Master thesis*, University of Washington, 2008.

[13] M. Dahl and I. Claesson and B. Sällberg and H. Akesson, A mixed analog-digital hybrid for speech enhancement purposes, *Proc. IEEE International Symposium on Circuits and Systems*, Kobe , Japan, 2005, 852- 855.

[14] M. Dahl and I. Claesson and B. Sällberg and H. Akesson, A mixed analog-digital hybrid for speech enhancement purposes, *Proc. IEEE International Symposium on Circuits and Systems*, Kobe, Japan, 2005, 852- 855.

[15] B. Sällberg, M. Dahl, Speech Enhancement implementations in the digital, analog and hybrid domain, *Proc. Swedish System on Chip Conference*, Stockholm, Sweden, 2005.

[16] S. Shamma, Encoding sound timbre in the auditory system, *IETE Journal of research*, 49(2), 2003, 193-205.

[17] S.M. Schimmel, Theory of modulation frequency analysis with applications to hearing devices, *Ph.D. dissertation*, University of Washington, 2007.

[18] P. Clark and L. E. Atlas, Time-frequency coherent modulation filtering of non-stationary signals, *IEEE transaction on Signal Processing*, 57(11), 2009, 4323-4332.

Enhancement of Spanish Oesophageal Speech Vowels using Coherent Subband modulator Kalman Filtering

Rizwan Ishaq [a,*], Begoña García Zapirain [a]
[a] *Desutotech-LIFE, University of Deusto, Bilbao, Spain*

Abstract. This paper proposes an Oesophageal Speech (OES) enhancement method, based on Kalman filtering. The Kalman filter is applied to modulators of OES frequency subbands instead of the fullband signal. The OES frequency subbands are decomposed into modulators and carriers components using coherent demodulation. In comparison with fullband Kalman filtering and pole stabilization, the proposed technique shows better results. The system performance is evaluated objectively and subjectively using the Harmonic to Noise Ratio (HNR) and Mean Opinion Score (MOS) respectively. Results have shown that Kalman filter in subband modulators processing is robust and efficient, improving the HNR by 4 to 5 dB for all Spanish vowels.

Keywords: Alaryngeal speech, Kalman filtering, filterbank, synthesis/analysis, modulation frequency domain

1. Introduction

The loss of speech production after total laryngectomy (advanced stage laryngeal cancer treatment) is one of extreme consequences for the larygectomee (patient). To regain the ability to produce speech, alternate means of speech production are needed. In the literature, there are three speech production methods currently available: Oesophageal Speech (OES), Treach-Oesophageal Speech (TES) and Electrolarynx (EL). OES and TES use esophagus as a voicing excitation to vocal tract, while EL uses an external vibrating device as voicing source to vocal tract. The OES is the preferred speech production because it does not require surgery (TES) or external devices (EL). But the OES has very low intelligibility due to irregular vibration of voicing source and altered vocal tract shape. The signal processing algorithms can be used to improve the intelligibility of OES.

In the literature some researchers have used signal processing algorithms for OES enhancement. The source-filter speech production model [1] has been used to decompose the OES into its source and filter components, and the source subsequently was replaced with the Liljencrants Fant (LF) glottal flow model for enhancing OES quality [2]. The modification to [2] has been obtained by processing fundamental frequency for intelligible OES [3]. The filter formants modification has provided significant improvement

*Corresponding author. E-mail: rizwanishaq@deusto.es

in intelligibility [4,5,6,7,8,9]. Statistical methods have shown significant results in improvement using OES to normal speech transformation [10].

The Kalman filter is an extensively used speech enhancement method for normal speech such as full-band Kalman filtering [11,12,13,14], and subband Kalman filtering [15,16,17,18,19,20,21,22,15,14,23, 24,25,26]. Speech production model inheritance and non-stationary processing are the advantages of Kalman filtering over other speech enhancement methods[16,19]. The Kalman filter has also been used for OES speech enhancement, demonstrating significant improvement. The first use of Kalman filtering for OES has provided significant enhancement[27]. The modification to [27] was introduced using poles stabilization, as shown in Figure. 1, and its results showed significantly improved speech quality [28,29,30,31]. To the best of our knowledge, up to now, nobody else has used Kalman filtering for OES enhancement.

The speech research community has been using the modulation domain for the last decade for speech enhancement , recognition, separation etc [32,33,34,35]. The modulation domain states that speech frequency subbands can be modeled as low frequency modulators and high frequency carriers [32,35]:

$$x_k(n) = m_k(n)c_k(n) \tag{1}$$

where $x_k(n)$, $m_k(n)$ and $c_k(n)$ are frequency subband k signal, its modulator and carrier respectively. It has been shown that modulators of speech signals are more important for intelligibility than carriers i.e. when modulators are replaced by some constant, speech is not intelligible [32,36,35].

This paper investigates the Kalman filter for OES enhancement in the modulation domain. The modulators $m_k(n)$ and carriers $c_k(n)$ are estimated using coherent demodulation [32,35] and Kalman filter is applied to the modulators without altering the carriers. The system was tested with the Harmonic to Noise (HNR) objectively and Mean Opinion Score (MOS) subjectively, for the Spanish male speaker OES vowels $\backslash a \backslash$, $\backslash e \backslash$,$\backslash i \backslash$, $\backslash o \backslash$, and $\backslash u \backslash$. The proposed system was compared with a reference system, Kalman Filtered Enhanced Speech (KF-ES), presented in [28,29], shown in Figure 1.

The structure of this paper is as follows. In Section 2 we describe the proposed method. The system components demodulation and frequency sub-band modulator Kalman Filter are described in Section 3 and 4 respectively. The optimal parameter estimation for optimal Kalman filtering is provided in Section 4.1. Section 5 explains the poles modification followed by a summary of the results and conclusion in Section 6 and 7.

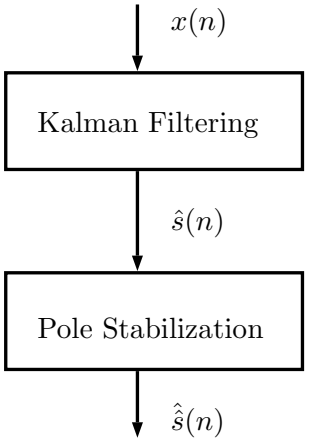

Fig. 1. Kalman Filtered Enhanced Speech (KF-ES) system ([28,29]).

2. Method

This section outlines the proposed method, Coherently Modulated Kalman Filtered Enhanced Speech (C-MKF-ES), also shown in Figure. 2. The first step in processing OES is to decompose the broadband OES into narrowband frequency subbands. The perfectly reconstructed filterbank of uniformally-spaced K bandpass filter, each having an impulse response of $h_k(n)$ using short-time Fourier transform (STFT), is used to decompose the speech signals into K narrowband subbands [36,35]. Mathematically:

$$x_k(n) = x(n) * h_k(n) \tag{2}$$

where $*$ is convolution operator.

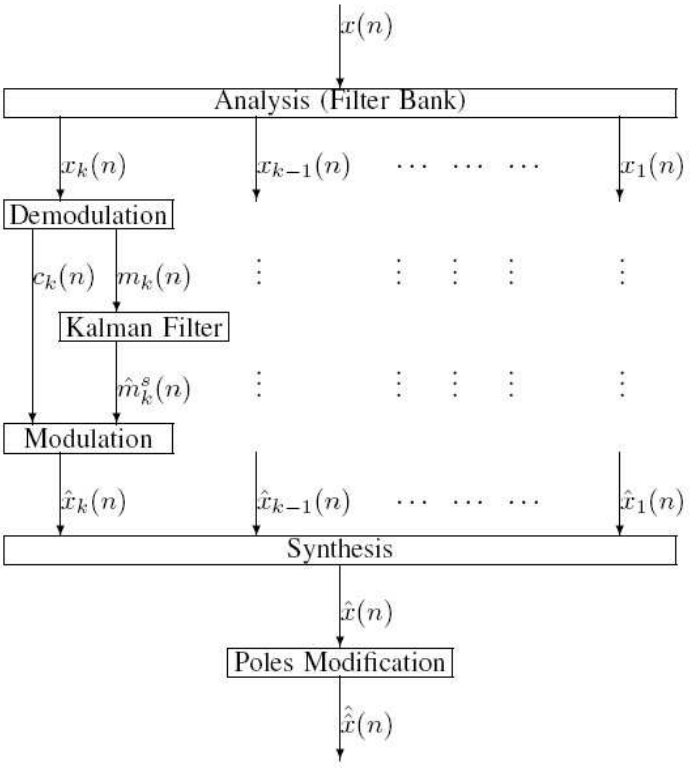

Fig. 2. Coherently Modulated Kalman Filtered Enhanced Speech (C-MKF-ES) system.

Each frequency subband is demodulated into carrier c_k and modulator m_k using coherent demodulation (Section. 3), mathematically:

$$x_k(n) = m_k(n)c_k(n) \tag{3}$$

The Kalman filter is applied to modulators $m_k(n)$ for modification (Section. 4). The modified frequency sub-band modulators $\hat{m}_k(n)$ are then modulated with carriers according to following equation:

$$\hat{x}_k(n) = \hat{m}_k(n)c_k(n) \tag{4}$$

where $\hat{x}_k(n)$ is the modified k frequency sub-band.

The synthesis filter bank transforms frequency subbands into a full-band signal:

$$\hat{x}(n) = \sum_{k=1}^{K} \hat{x}_k(n) \tag{5}$$

The final enhanced version $\hat{\hat{x}}(n)$ is obtained by passing $\hat{x}(n)$ through a poles modification process (Section. 5).

3. Demodulation

The demodulation is a process of estimating the frequency subband modulators $m_k(n)$ and carriers $c_k(n)$. The demodulation can be coherent or non-coherent. The non-coherent demodulation estimates $m_k(n)$ and $c_k(n)$ separately [32], while coherent demodulation (paper used coherent demodulation) of $m_k(n)$ depends on the estimation of $c_k(n)$ [36,35].

3.1. Coherent Demodulation

The coherent demodulation estimates the modulator in terms of an explicitly estimated carrier signal [32,35], mathematically:

$$m_k(n) = x_k(n)c_k^*(n) \tag{6}$$

where $c_k^*(n)$ is the carrier and given as:

$$c_k^*(n) = e^{-j\phi(n)} \tag{7}$$

where phase $\phi_k(n)$ is:

$$\phi_k(n) = \sum_{p=o}^{n} \omega_k(p) \tag{8}$$

where $\omega_k(p)$ is the instantaneous frequency of subband k and is defined according to the Center-of-Gravity (CoG) approach as the average frequency of instantaneous spectrum of $x_k(n)$ [32,36,35]:

$$\omega_k(n) = \frac{\sum_{i=0}^{L-1} \alpha(i)|X_k(i,n)|^2}{\sum_{i=0}^{L-1} |X_k(i,n)|^2} \tag{9}$$

where L is Discrete Fourier Transform (DFT) length, $\alpha(i)$ is the weighting function:

$$\alpha(i) = \begin{cases} 2\pi i/L & 0 \leq i \leq L/2 \\ 2\pi i/L - 2\pi & L/2 < i < L \end{cases} \tag{10}$$

and $X_k(i,n)$ is the subband Fourier transform, and given as:

$$X_k(i,n) = \sum_p w(p)x_k(n+p)e^{-j2\pi(i/L)p}, i = 0 : L-1 \tag{11}$$

where $w(p)$ is a window function, i.e. Hamming or Hanning window.

There are other type of demodulation available in literature i.e. convex optimized demodulation [37] and probabilistic amplitude demodulation [38], but they are computationally very time consuming.

4. Coherent Subband Modulator Kalman Filter

The modulators $m_k^s(n)$ of speech signal can be modeled as AR process, and represented by the following linear equation:

$$m_k^s(n) = \sum_{j=1}^{p} a_{k,j}^m m_k^s(n-j) + \omega_k^s(n) \tag{12}$$

where $a_{k,j}^m$ and p are the prediction coefficients and order. ω_k^s is a white Gaussian process with zero mean and variance $\sigma_{\omega_k^s}^2$. The noise modulator $m_k^v(n)$ is also an AR process:

$$m_k^v(n) = \sum_{j=1}^{q} b_{k,j}^m m_k^v(n-j) + \zeta_k^v(n) \tag{13}$$

where $b_{k,j}^m$ are prediction coefficients, q prediction order, and ζ_k^v white Gaussian processing with zero mean and variance $\sigma_{\zeta_k^v}^2$ The state-space domain representation for Kalman filtering usage is:

$$\bar{m}_k^s(n) = F_k \bar{m}_k^s(n) + G_k \bar{\omega}_k^s(n) \tag{14}$$

$$m_k(n) = H_k^T \bar{m}_k^s(n) \tag{15}$$

where

$$F_k = \begin{bmatrix} \begin{bmatrix} 0 & 1 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \\ -a_{k,p}^m & -a_{k,p-1}^m & \dots & -a_{k,1}^m \end{bmatrix} & 0 \\ 0 & \begin{bmatrix} 0 & 1 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \\ -b_{k,q}^m & -b_{k,q-1}^m & \dots & -b_{k,1}^m \end{bmatrix} \end{bmatrix}$$

$$G_k = \begin{bmatrix} [1,0,\dots,0]_{p\times 1} & 0 \\ 0 & [1,0,\dots,0]_{q\times 1} \end{bmatrix}$$

$$H_k^T = \left[{}_{[1,0,\ldots,0]_{p\times 1}^T} \; {}_{[1,0,\ldots,0]_{q\times 1}^T} \right]$$

$$\bar{\omega_k^s}(n) = \begin{bmatrix} \omega_k^s(n) \\ \zeta_k^v(n) \end{bmatrix}$$

$$\bar{m}_k^s(n) = [m_k^s(n-p+1), \ldots, m_k^s(n), m_k^v(n-q+1), \ldots, m_k^v(n)]^T$$

The covariance matrix for $\bar{\omega_k^s}(n)$ is described by following mathematical relation:

$$Q_k = E[\bar{\omega_k^s}(n)\bar{\omega}_k^s(n)^T] = \begin{bmatrix} \sigma^2_{\omega_k^s} & 0 \\ 0 & \sigma^2_{\zeta_k^v} \end{bmatrix} \tag{16}$$

The Kalman filter estimates the $m_k^s(n)$, providing observation vector $\{m_k(1), m_k(2), ...m_k(n)\}$ [39]:

$$\hat{m}_k^s(n) = F_k\hat{m}_k^s(n-1) + K_k(n)[m_k(n) - H_k^T F_k \hat{m}_k^s(n-1)] \tag{17}$$

where $K_k(n)$ is the Kalman gain and given as:

$$K_k(n) = \frac{P_k(n|n-1)H_k}{[H_k^T P_k(n|n-1)H_k]} \tag{18}$$

$$P_k(n|n-1) = F_k P_k(n-1)F^T + G_k Q_k G_k^T \tag{19}$$

$$P_k(n) = [I - K_k(n)h^T]P_k(n|n-1) \tag{20}$$

where $P_K(n)$ and $P_k(n|n-1)$ are the filtering and prediction-error covariance matrices respectively. At time instant n, the speech sample is described by,

$$\hat{m}_k^s(n) = H_k^T \hat{m}_k^s(n) \tag{21}$$

4.1. Parameter Estimation

The optimum results of Kalman filtering can be obtained when the estimation of AR coefficients $a_{k,j}^m, b_{k,j}^m$ and variances $\sigma^2_{\omega_k^s}, \sigma^2_{\zeta_k^v}$ are optimal. Poor estimation of these parameters resulted in distorted speech. This section provides the optimal estimation of $a_{k,j}^m, \sigma^2_{\omega_k^s}$ utilizing Weight Linear Prediction (WLP) and $b_{k,j}^m, \sigma^2_{\zeta_k^v}$ using Linear Prediction (LP).

4.1.1. Weighted Linear Prediction (WLP)

The conventional Linear Prediction (LP) estimates the AR coefficients by minimizing the error between estimated and measured signals [1].

$$\hat{m}_k(n) = \sum_{j=1}^{p} a_{k,j}^m m_k(n-j) \tag{22}$$

The error signal $\epsilon_k(n)$ is:

$$\epsilon_k(n) = \sum_{n=1}^{N} [m_k(n) - \sum_{j=1}^{p} a_{k,j}^m m_k(n-j)] \tag{23}$$

The AR coefficients $a_{k,j}^m$ are estimated using minimum square error criterion,

$$\epsilon_k^2(n) = \sum_{n=1}^{N} (m_k(n) - \sum_{j=1}^{p} a_{k,j}^m m_k(n-j))^2 \tag{24}$$

The minimum of the above equation can be obtained by setting its derivative with respect to $a_{k,j}^m$ zero:

$$\frac{\partial \epsilon_k^2}{\partial a_{k,i}^m} = \sum_{n=1}^{N} (2(m_k(n) - \sum_{j=1}^{p} a_{k,j}^m m_k(n-j)) m_k(n-i)) = 0 \tag{25}$$

$$= -2 \sum_{n=1}^{N} m_k(n) m_k(n-i) + 2 \sum_{n=1}^{N} \sum_{j=1}^{p} a_{k,j}^m m_k(n-j) m_k(n-i)$$

$$= 0, \forall i = 1, 2, \ldots, p \tag{26}$$

$$\sum_{n=1}^{N} m_k(n) m_k(n-i) = \sum_{j=1}^{p} a_{k,j}^m \sum_{n=1}^{N} m_k(n-j) m_k(n-i) \tag{27}$$

The covariance function is defined as:

$$r_k^{mm}(i,j) = \sum_{n=1}^{N} m_k(n-i) m_k(n-j)$$

which results in:

$$r_k^{mm}(i,0) = \sum_{j=1}^{p} r_k^{mm}(i,j) a_{k,j}^m \tag{28}$$

In the matrix form:

$$\begin{bmatrix} r_k^{mm}(1,0) \\ r_k^{mm}(2,0) \\ \vdots \\ r_k^{mm}(p,0) \end{bmatrix} = \begin{bmatrix} r_k^{mm}(1,1) & r_k^{mm}(1,2) & \dots & r_k^{mm}(1,p) \\ r_k^{mm}(2,1) & r_k^{mm}(2,2) & \dots & r_k^{mm}(2,p) \\ \vdots & \vdots & \dots & \vdots \\ r_k^{mm}(p,1) & r_k^{mm}(p,2) & \dots & r_k^{mm}(p,p) \end{bmatrix} \begin{bmatrix} a_{k,1}^m \\ a_{k,2}^m \\ \vdots \\ a_{k,p}^m \end{bmatrix} \tag{29}$$

in compact form

$$\vec{r} = R\vec{a} \tag{30}$$

and solution for $\vec{a}$ given by,

$$\vec{a} = R^{-1}\vec{r} \tag{31}$$

The predicted coefficients are degraded due to conventional LP sensitivity to background noise [40,41]. The weighting function is introduced to overcome conventional LP problems [40,41]. The weighting function is calculated using the Short-Time Energy (STE) $\Psi_k(n)$[40]:

$$\Psi_k(n) = \sum_{j=1}^{M} m_k^2(n-j) \tag{32}$$

where M is number of samples, used to estimate energy. The prediction error by introducing the weighting function is:

$$\epsilon_k^2(n) = \sum_{n=1}^{N} [m_k(n) - \sum_{j=1}^{p} a_{k,j}^m m_k(n-j)]^2 \Psi_k(n) \tag{33}$$

Solving for $a_{k,j}^m$, we have:

$$\begin{bmatrix} a_{k,1}^m \\ a_{k,2}^m \\ \vdots \\ a_{k,p}^m \end{bmatrix} = \begin{bmatrix} r_k^{mm}(1,1) & r_k^{mm}(1,2) & \dots & r_k^{mm}(1,p) \\ r_k^{mm}(2,1) & r_k^{mm}(2,2) & \dots & r_k^{mm}(2,p) \\ \vdots & \vdots & \dots & \vdots \\ r_k^{mm}(p,1) & r_k^{mm}(p,2) & \dots & r_k^{mm}(p,p) \end{bmatrix}^{-1} \begin{bmatrix} r_k^{mm}(1,0) \\ r_k^{mm}(2,0) \\ \vdots \\ r_k^{mm}(p,0) \end{bmatrix} \tag{34}$$

where r_k^{mm} is the covariance of the modulators and given as [23]:

$$r_k^{mm}(i,j) = \sum_{n=1}^{N} \Psi_k(n) m_k(n-i) m_k(n-j) \tag{35}$$

The variance $\sigma_{\omega_k^s}^2$ is calculated according to following relation:

$$\sigma_{\omega_k^s}^2 = r_k^{mm}(0,0) - \sum_{i=1}^{p} a_{k,i}^m r_k^{mm}(0,i) \tag{36}$$

4.1.2. Noise parameters

The noise signal modulators are assumed to be known, and the AR coefficients and variance can be estimated using the covariance function $r_k^{m_v m_v}(i,j)$:

$$r_k^{m_v m_v}(i,j) = \sum_{n=1}^{N} m_k^v(n-i) m_k^v(n-j) \tag{37}$$

Solving for $b_{k,i}^m$, we have:

$$\sum_{i=1}^{q} b_{k,i}^m r_k^{m_v m_v}(j,i) = r_k^{m_v m_v}(j,0), \forall j = 1,2,\ldots,q \tag{38}$$

$$\begin{bmatrix} b_{k,1}^m \\ b_{k,2}^m \\ \vdots \\ b_{k,q}^m \end{bmatrix} = \begin{bmatrix} r_k^{m_v m_v}(1,1) & r_k^{m_v m_v}(1,2) & \ldots & r_k^{m_v m_v}(1,q) \\ r_k^{m_v m_v}(2,1) & r_k^{m_v m_v}(2,2) & \ldots & r_k^{m_v m_v}(2,q) \\ \vdots & \vdots & \vdots & \vdots \\ r_k^{m_v m_v}(q,1) & r_k^{m_v m_v}(q,2) & \ldots & r_k^{m_v m_v}(q,q) \end{bmatrix}^{-1} \begin{bmatrix} r_k^{m_v m_v}(1,0) \\ r_k^{m_v m_v}(2,0) \\ \vdots \\ r_k^{m_v m_v}(q,0) \end{bmatrix} \tag{39}$$

The variance $\sigma_{\zeta_k^v}^2$ according to [23]:

$$\sigma_{\zeta_k^v}^2 = r_k^{m_v m_v}(0,0) - \sum_{i=1}^{q} b_{k,i}^m r_k^{m_v m_v}(0,i) \tag{40}$$

5. Poles modification

The enhanced speech signal $\hat{x}(n)$ is further enhanced by modifying the AR coefficients (poles) using the Line Spectral Frequency (LSF) pairs [42], following these steps:

- estimating AR coefficients for short segment of speech
- converting AR coefficients to LSF pairs
- modifying LSF according to the following equation:

$$\hat{lsf}_i = lsf_{i-1} + d_{i-1} + \frac{d_{i-1}^2}{d_{i-1}^2 + d_i^2}((lsf_{i+1} - lsf_{i-1}) - (d_i + d_{i-1})) \tag{41}$$

  where lsf_i and $\hat{lsf}_i$ are the original and modified LSF of a frame respectively.

$$d_i = \alpha(lsf_{i+1} - lsf_i), i = 2,\ldots,p-1 \tag{42}$$

  where p is the prediction order, and α the controlling constant for modification.
- converting back the modified LSF pairs to AR coefficients

6. Results and Discussion

The system uses the Spanish OES vowels {a,e,i,o,u}. Six male OES subjects of the speech rehabilitation center (there are no female subjects in the center) participated in vowel recording. Each vowel was uttered four times. The recording sampling frequency was 44100 Hz and down sampled to 16 KHz for computation efficiency. The system tuning parameters are given in Table 1. The system was evaluated objectively using Harmonic to Noise Ratio (HNR) [43,44], and subjectively through Mean Opinion Score (MOS). The VoiceSauce matlab based speech analysis toolbox was used for HNR calculation [45].

Table 1
The System parameters setting for simulations

Fullband Kalman Filter (KF-ES)	**Coherent demodulated Kalman Filter (C-MKF-ES)**
frame size: 40 ms , overlap: 15 ms	frame size: 40 ms , overlap: 15 ms
(p,q): (16,16), (16,8), (12,8), (8,4), (4,2)	(p,q): (16,16), (16,8), (12,8), (8,4), (4,2)
Poles stabilization [27,28,29]	**Poles Modification [42]**
Modulus threshold $\tau_M = 0.001$	prediction order p: 16
phase threshold $\tau_\theta = \pi/8$	$\alpha = 0.3, 0.4, 0.5$
modulus stabilization constant $\kappa_M = 0.5$	**Filterbank setting**
phase stabilization constant $\kappa_\theta = 0.001$	Uniformly-spaced 16 channel filterbank [32,36]

6.1. Harmonic to Noise Ratio (HNR)

The HNR is an objective measurement parameter for assessing the noise level in human voice signals [43]. The average mean HNR for different vowels in table 2 has shown improvement of around 6 dB for the proposed system. The noise and speech prediction order plays an important role in enhancement. The poles modification sharpen the vocal tracts formants, which provide enhancement as can be seen in table 2.

Table 2
Average Mean Harmonic to Noise ratio improvement for Spanish OES vowels with/without different α values

Method$_{\downarrow}$	$(p,q)_{\rightarrow}$	(4,2)	(8,4)	(12,8)	(16,8)	(16,16)	$\alpha_{\downarrow}$
KF-ES (dB)		1.23	2.33	2.58	2.88	3.01	-
		3.90	3.87	4.01	4.88	5.11	-
		5.47	5.31	6.10	7.09	7.59	0.3
C-MKF-ES (dB)		4.51	4.73	5.44	6.12	7.19	0.4
		4.97	4.60	5.12	6.10	6.89	0.5
		$\backslash a \backslash$					
KF-ES (dB)		1.41	2.13	2.61	2.91	3.41	-
		2.19	2.87	3.15	3.51	4.12	-
		4.13	4.18	4.90	5.01	5.88	0.3
C-MKF-ES (dB)		3.93	4.81	4.97	5.32	5.86	0.4
		4.03	4.33	4.87	5.23	5.79	0.5
		$\backslash e \backslash$					
KF-ES (dB)		1.89	2.11	2.94	3.13	3.60	-
		1.87	2.13	2.79	3.07	3.89	-
		2.41	2.97	3.89	4.08	5.17	0.3
C-MKF-ES (dB)		3.13	3.71	4.17	5.12	6.32	0.4
		2.67	3.41	4.13	5.42	6.21	0.5
		$\backslash i \backslash$					
KF-ES (dB)		2.13	2.53	3.03	3.44	2.67	-
		2.01	2.87	3.14	3.65	4.23	-
		2.87	3.56	5.32	5.39	5.98	0.3
C-MKF-ES (dB)		2.96	3.76	4.01	4.78	5.18	0.4
		2.79	3.18	4.32	5.31	6.13	0.5
		$\backslash o \backslash$					
KF-ES (dB)		0.98	1.23	2.01	2.67	2.88	-
		1.75	2.33	3.21	3.79	4.63	-
		2.86	2.90	3.45	3.96	4.70	0.3
C-MKF-ES (dB)		1.89	2.45	3.87	4.88	6.01	0.4
		2.03	3.10	3.21	2.77	3.02	0.5
		$\backslash u \backslash$					

6.2. Mean Opinion Score (MOS)

The system was tested using MOS (ranging from 1, bad, to 5, excellent) with ten listeners. None of them had listened to OES before. Most of the listeners were unable to understand non-processed samples, and gave minimum MOS scores as can be seen in table 3. The average MOS for each vowel is shown in table 3. It can be observed that the proposed system outperforms the KF-ES system significantly. The poles modification has provided much better enhancement as the average MOS is above 3.

Table 3
Average Mean Opinion Score (MOS) for Spanish OES vowels with/without different α values

Method$_{\downarrow}$ $(p,q)\rightarrow$	(4,2)	(8,4)	(12,8)	(16,8)	(16,16)	$\alpha_{\downarrow}$
Original	1.04	1.04	1.04	1.04	1.04	-
KF-ES	1.89	2.23	2.50	2.63	2.21	-
	2.03	2.31	2.33	2.56	2.67	-
	2.76	2.89	3.01	3.03	2.99	0.3
C-MKF-ES	2.65	2.97	3.11	3.14	3.21	0.4
	2.43	3.13	3.16	3.23	3.25	0.5
		$\backslash a \backslash$				
Original	1.07	1.07	1.07	1.07	1.07	-
KF-ES	1.76	1.83	2.01	2.32	2.21	-
	2.01	2.05	2.09	2.08	2.11	-
	2.21	2.24	2.28	2.31	2.35	0.3
C-MKF-ES	2.26	2.29	2.45	2.49	2.67	0.4
	2.34	2.87	2.95	3.01	3.13	0.5
		$\backslash e \backslash$				
Original	1.0	1.0	1.0	1.0	1.0	-
KF-ES	1.61	1.82	1.86	1.92	1.97	-
	1.79	1.94	2.03	2.05	2.07	-
	2.01	2.23	2.56	2.04	2.34	0.3
C-MKF-ES	2.09	2.34	2.37	2.45	2.87	0.4
	2.11	2.23	2.42	2.80	3.03	0.5
		$\backslash i \backslash$				
Original	1.10	1.10	1.10	1.10	1.10	-
KF-ES	1.81	1.84	1.82	1.86	1.91	-
	2.12	2.23	2.26	2.24	2.56	-
	2.09	2.35	2.67	2.69	2.71	0.3
C-MKF-ES	2.07	2.45	2.49	2.51	2.89	0.4
	2.11	2.38	2.87	3.03	3.23	0.5
		$\backslash o \backslash$				
Original	1.0	1.0	1.0	1.0	1.0	-
KF-ES	1.67	1.89	1.92	1.96	2.01	-
	1.89	1.81	1.73	1.82	1.98	-
	1.65	1.93	2.03	2.32	2.84	0.3
C-MKF-ES	1.84	2.03	2.08	2.34	2.56	0.4
	1.54	2.34	2.42	2.59	3.32	0.5
		$\backslash u \backslash$				

The spectrogram of OES vowel /a/ and the enhanced signal are shown in Fig. 3 and 4 respectively. The enhanced version OES vowel /a/ spectrogram shows formants very clearly compared with the spectrogram of the non-processed OES vowel /a/.

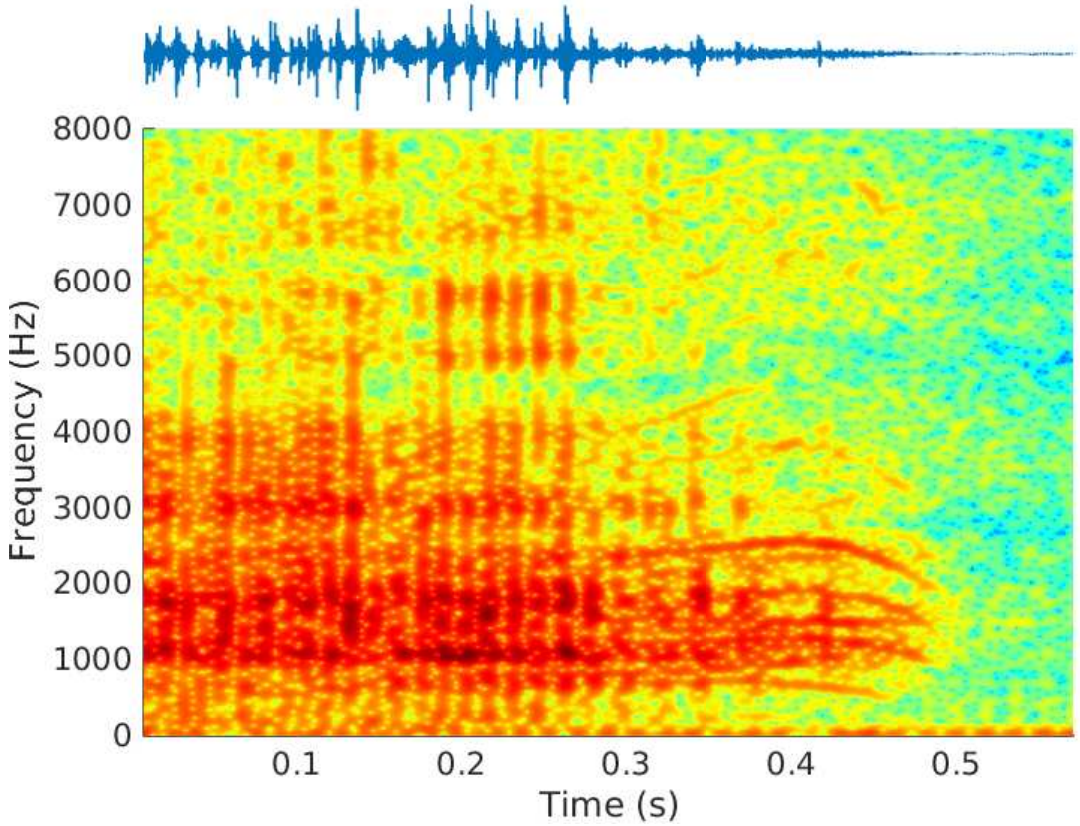

Fig. 3. OES vowel /a/ Spectrogram and waveform

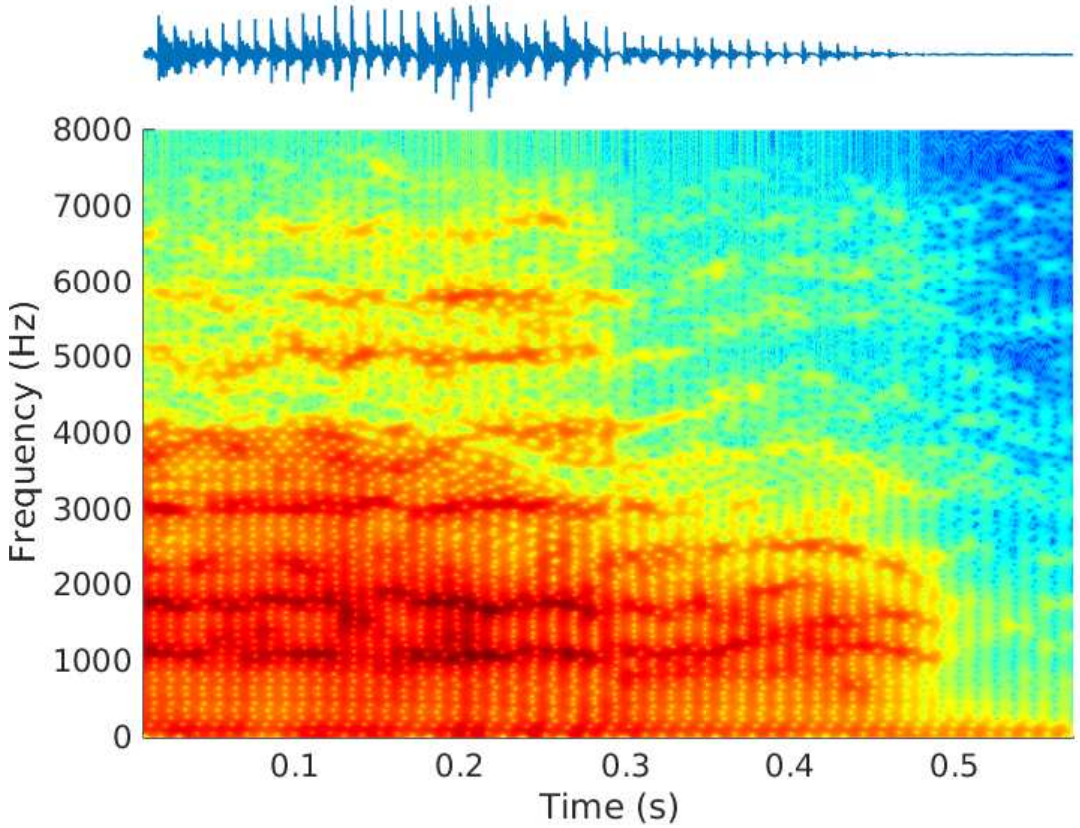

Fig. 4. Enhanced OES vowel /a/ Spectrogram and waveform

7. Conclusion

In this paper a new method is presented for enhancing Oesophageal Speech (OES). Kalman Filtering was applied to modulators of frequency subbands instead of the full band signal. The frequency subband modulators are estimated coherently. The Autoregressive (AR) and noise parameters for speech and noise modulators are estimated using Weighted Linear Predication (WLP) instead of Linear Prediction (LP). The AR parameters are further modified using a poles modification method. The Harmonic to Noise Ratio (HNR) and Mean Opinion Score (MOS) are used to assess the system's capability with Spanish male OES vowels $\backslash a \backslash$, $\backslash e \backslash$, $\backslash i \backslash$, $\backslash o \backslash$ and $\backslash u \backslash$. The results showed that the proposed system, the coherent modulator Kalman Filter (C-MKF-ES), outperforms the fullband Kalman filter (KE-ES) subjectively as well objectively. The system has only been evaluated with male subjects due to non-availability of female OES subjects.

References

[1] Makhoul J. Linear Prediction: a tutorial review. 1975;63(4):561–580.

[2] Yingyong Q, Bernd W, Ning B. Enhancement of Female Esophageal and Tracheoesophageal Speech [article]. Acoustical Society of America. 1995;98(5, Pt1):2461–2465.

[3] Qi Y. Replacing Tracheoesophageal Voicing Source using LPC Synthesis [article]. Acoustical Society of America. 1995;5:1228–1235.

[4] Ali RH, Jebara SB. Esophageal Speech Enhancement Using Excitation Source Synthesis and Formant Structure Modification [article]. SITIS. 2006;p. 615–624.

[5] Sirichokswad R, Boonpramuk P, Kasemkosin N, Chanyagorn P, Charoensuk W, Szu HH. Improvement of Esophageal Speech using LPC and LF Model [article]. Internation Conf on Biomedical and Pharamaceutical Engineering 2006. 2006;p. 405–408.

[6] Tull RG, Rutledge JC. Linear Predictive Synthesis of Vowels for Pitch Enhancement of Female Geriatric Esophageal Speech. In: Engineering in Medicine and Biology Society, 1993. Proceedings of the 15th Annual International Conference of the IEEE; 1993. p. 1359–1360.

[7] Prosek RA, Vreeland LL. The Intelligibility of Time Domain Edited Esophageal Speech [article]. American Speech Language Hearing Association. 2001;44:525–534.

[8] Kenji M, Noriyo H, Noriko K, Hajime H. Enhancement of Esophageal Speech using Formant Synthesis [article]. Acoustic Sci and Tech. 2002;p. 69–76.

[9] Kenji M, Noriyo H. Enhancement of Esophageal Speech using Formant Synthesis [article]. Acoustics, Speech and Signal Processing, International conf. 1999;p. 81–85.

[10] Doi H, Nakamura K, Toda T, Saruwatari H, Shikano K. Statistical Approach to Enhancing Esophageal Speech Based on Gaussian Mixture Models. In: Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on; 2010. p. 4250–4253.

[11] Paliwal KK, Basu A. A speech enhancemet method based on Kalman Filtering. IEEE Int Conf Acoust, Speech, Signal Processing. 1987;12:177–180.

[12] Popescu DC, Zeljkovic I. Kalman filtering of colored noise for speech enhancement. In: Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on. vol. 2; 1998. p. 997–1000.

[13] Puder H. Kalman-Filter in subbands for noise reduction with enhanced pitch-adaptive speech model estimation. In: European Transactions on Telecommunications. vol. 13; 2002. p. 139–148.

[14] So S, Paliwal KK. Fast Converging Iterative Kalman Filtering for Speech Enhancement using Long and Overlapped Tapered Windows with large Side Lobe Attenuation. In: Interspeech(ISCA)-2010, Makuhari, Chiba, Japan; 2010. p. 1081–1084.

[15] Leandro Aureliano de S, Macelo Basalio J. Noise reduction in biomedical speech signal processing based on time and frequency Kalman filtering combined with spectral subtraction. Computers & Electrical Engineering. 2008;34:154–164.

[16] So S, Paliwal KK. Modulation-domain Kalman filtering for single-channel speech enhancement. Speech Commun. 2011 Jul;53(6):818–829.

[17] Gabrea M, Grivel E, Najim M. A Single Microphone Kalman Filter- Based Noise Canceller. IEEE Signal processing Letters. 1999;6(3):55–57.

[18] Le PN, Ambikairajah E. Non-Uniform Sub-Band Kalman Filtering for Speech Enhancement. Internation Conference on Signal Processing and Communication Systems (ICSPCS). 2007;.

[19] Wu WR, Chen PC. Subband Kalman Filtering for Speech Enhancement. Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on. 1998;45(8):1072–1083.
[20] Huai YC, Ngee KS, Susanto R. Subband Kalman filtering incorporating masking properties for noisy speech signal. Speech Commun. 2007 Jul;49(7-8):558–573.
[21] So S, Paliwal KK. Supressing the influence of additive noise on the Kalman gain for low residual noise speech enhancement. Elsever, Speech communication 53. 2011;53:355–378.
[22] So S, Paliwal KK. A long state vector Kalman Filter for Speech Enhancement. In: Proceedings of the 9th Annual Conference of the International Speech Communication Association; 2008. p. 391–394.
[23] Wen Rong W, Po Cheng C, HwaiTsu C, Chun Hung K. Frame-based subband Kalman filtering for speech Enhancement. In: Signal Processing Proceedings, 1998. ICSP' 98. International Conference. vol. 1; 1998. p. 682–685.
[24] Wu WR, Chen PC. Frame-based Sub-Band Kalman Filtering for Speech enhancement. Acoustic Society of America. 2003;113.
[25] Weixiu D, Driessen P. Speech Enhancement Based on Kalman Filtering and EM Algorithm. IEEE Pacific Rim Conf on Communication, Computers and Signal Processing, 1991. 1991;p. 142–145.
[26] Sorqvist P, Handel P, Ottersten B. Kalman filtering for low distortion speech enhancement in mobile communication. In: Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on. vol. 2; 1997. p. 1219–1222.
[27] Garcia B, Ruiz I, Vicente J, Alonso A. Formants Measurement for Esophageal Speech using Wavelet with Band and Resoultion Adjustment [article]. IEEE Symposium on Signal Processing and Information Technology. 2006;p. 320–325.
[28] Garcia B, Mendez A. Oesophageal Speech Enhancement using poles stablization and Kalman Filtering [article]. ICASSP. 2008;p. 1597–1600.
[29] Ibon OR, Garcia B, Amaia ZM. New Approach for Oesophageal Speech Enhancement [article]. 10th International conference, ISSPA. 2010;5:225–228.
[30] Ishaq R, Zapirain BG, Shahid M, Lovstrom B. Subband Modulator Kalman Filtering for Signla channel Speech Enhancement. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2013. p. 7442–7446.
[31] Ishaq R, Zapirain BG. Optimal subband Kalman filter for normal and oesophageal speech enhancement. Bio-Medical Materials and Engineering. 2014;24:3569–3578.
[32] Clark CP. Effective Coherent Modulation Filtering and Interpolation of Long Gaps in Acoustic Signals. University of Washington; 2008.
[33] Ishaq R. Adaptive Gain Equalizer and Modulation Frequency Domain for Noise Reduction [Master Thesis]. School of Engineering. Blekinge Institute of Technology; 2010.
[34] Shahid M, Ishaq R, Sallberg B, Grbic N, Lovstrom B, Claesson I. Modulation Domain Adaptive Gain Equalizer for Speech Enhancement. In: Signal and Image Processing Application 2011, by IASTED; 2011. .
[35] Schimmel S, Atlas L. Coherent envelope detection for modulation filtering of speech. In: IEEE International Conference on Acoustic, Speech and Signal Processing. vol. 1; 2005. p. 221–224.
[36] Atlas L, Clark P, Schimmel S. Modulation Toolbox Version 2.1 for Matlab; 2010. Available from: http://isdl.ee.washington.edu/projects/modulationtoolbox/.
[37] Boyd S, Vandenberghe L. Convex Optimization. 1st ed. Cambridge, UK: Cambridge University Press; 2004.
[38] Turner RE, Shani M. Probabilistic Amplitude Demodulation;.
[39] Gibson JD, Koo B, Gray SD. Filtering of Colored Noise for Speech Enhancement and Coding. IEEE Trans on Signal Processing. 1991;(8):1732–1742.
[40] Ma C, Kamp Y, Willems LF. Robust Single Selection for Linear Prediction analysis of voiced Speech. In: Speech Communication. vol. 2; 19983. p. 69–81.
[41] Magi C, Pohjalainen J, Bäckström T, Alku P. Stabilized Weighted Linear Prediction. In: Speech Communcation. vol. 5; 2009. p. 401–411.
[42] Raitio T, Sunni A, Pulakka H, Vainio M, Alku P. Comparison of Formant Enhancement Methods for HMM-Based Speech Synthesis; 2010. p. 334–339.
[43] Qi Y, Hillman RE. Temporal and Spectral Estimations of Harmonics-to-Noise Ratio in human voice signals. Acoustic Society of America. 1997 July;102(1):537–543.
[44] Sousa R, Ferriera A. Evaluation of existing Harmonics-to-Noise Ratio methods for voice assessment. In: Signal Processing Algorithms, Architectures, Arrangements, and Applications (SPA); 2008. p. 73–78.
[45] Alwan A. VoiceSauce: A Program for Voice Analysis @ONLINE; 2012. Available from: http://www.ee.ucla.edu/ spapl/voicesauce/.

Bibliography

[1] URL http://www.entnet.org/HealthInformation/laryngealCancer.cfm.

[2] URL http://publications.cancerresearchuk.org/downloads/Product/CS_KF_LARYNGEAL.pdf.

[3] G. Fant. Acoustic theory of speech production. Mouton, The Hauge, 1960.

[4] P. Alku. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. In *Speech communication*, volume 11, pages 109–118, 1992.

[5] Johan Sundberg. The acoustic of the signing voice. *Scientific American*, 236: 104–112, 1977.

[6] URL http://radonc.ucsd.edu/patient-info/Documents/WYNTK_larynx.pdf.

[7] URL http://emedicine.medscape.com/article/883689-overview.

[8] URL https://patienteducation.osumc.edu/Documents/esoph-sp.pdf.

[9] Dunn H., K. The calculation of vowel resonance, and an electrical vocal tract. *The Journal of the Acoustical Society of America*, 22:740–753, 1950.

[10] Gunnar Fant. Analysis and synthesis of speech processes. *Manual of Phonetics*, 2:173–277, 1968.

[11] Mark R. P. Thomas. *Glottal-Synchronous Speech Processing*. PhD thesis, Communication and Signal Processing Research Group, Department of Electrical Engineering, Imperial College London, 2010.

[12] Peter Birkholz. URL http://www.vocaltractlab.de/index.php?page=background-articulatory-synthesis.

[13] John I. Makhoul and Jared J. Wolf. Linear prediction and the spectral analysis of speech. Technical report, Advanced Research Projects Agency, Arlington, Virgina, 1972.

[14] J. Makhoul. Linear prediction: a tutorial review. 63(4):561–580, 1975.

[15] Carl Magi, Jouni Pohjalainen, Tom Backstrom, and Paavo Alku. Stabilised weighted linear prediction. *Speech Communication*, 51:401–411, 2009.

[16] Kamp Y Ma, C. and L. F. Willems. Robust signal selection for linear prrediction analysis of voiced speech. *Speech Communication*, 12:69–81, 1993.

[17] Kamp Y. and Willems. L. F. Ma, C. Robust single selection for linear prediction analysis of voiced speech. In *Speech Communication*, volume 2, pages 69–81, 19983.

[18] Tomi Kinnunen Jouni Pohjalainen, Rahim Saeidi and Paavo Alku. Extended weighted linear prediction (xlp) analysis of speech and its application to speaker verification in adverse conditions. In *INTERSPEECH*, 2010.

[19] El-Jaroudi and J. A.; Makhoul. Discrete all-pole modeling. *Signal Processing, IEEE Transactions on,*, 39(2):411–423, 1991.

[20] B. S. Atal and L. Hanauer, Suzanne. Speech analysis and synthesis by linear prediction of the speech wave. *The Journal of the Acoustical Society of America*, 50:637–655, 1971.

[21] J. D. Markel and Jr. A. H. Gray. Linear prediction of speech. *Springer-Verlag Berlin Heidelberg New York*, 12, 176.

[22] Gilles Degottex. *Glottal Source and Vocal-Tract Separation, Estimation of glottal parameters, voice transformation and synthesis using a glottal model.* PhD thesis, EcoleDoctorale Informatique, Telecommunications et Electronique (EDITE), 2010.

[23] A. E. Rosenberg. Effect of glottal pulse shape on the quality of natural vowels. *The Journal of the Acoustical Society of America*, 49:583–590, 1971.

[24] G. Fant. Vocal source analysis - a progress report. Technical Report 20, STL-QPSR 31-53, 1979.

[25] G. Fant. The lf model revisited. transformation and frequency domain analysis. *STL-QPSR*, pages 121–156, 1995.

[26] R. Veldhuis. A computationally efficient alternative for the liljencrants-fant model and its perceptual evaluation. *Journal of Acoustical society of America*, 103:566–571, 1998.

[27] Boris Doval, Christophe d'Alessandro, and Nathalia. The voice source as a causal/anticausal linear filter. In *In Proc. ISCA Voice Quality: Functions, Analysis and Synthesis (VOQUAL)*, 2003.

[28] H. Klatt Dennis and Laura C. Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87:820–857, 1990.

[29] H. Fujisaki and M. Ljungqvist. Proposal and evaluation of models for the glottal source waveform. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 11, pages 1605–1608, 1986.

[30] Backstrom T. Magi, C. Pohjalainen and P. Alku. Stablized weighted linear prediction. In *Speech Communcation*, volume 5, pages 401–411, 2009.

[31] R. Oppenheim, A. Schafer. Homomorphic analysis of speech. In *IEEE Transactions on Audio and Electroacoustics*, volume 16, pages 221–226, 1968.

[32] Monson H. Hayes. *Statistical Digital Signal Processing and Modeling*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1996. ISBN 0471594318.

[33] T. Drugman, B. Bozkurt, and T. Dutiot. Causal-anticausal decomposition of speech using complex cepstrum for glottal source estimation. *Speech Communication*, 53:855–866, 2011.

[34] T. Drugman, B. Bozkurt, and T. Dutiot. Complex cepstrum-based decomposition of speech for glottal source estimation. In *Interspeech*, 2009.

[35] T. Drugman and T. Dutiot. Chirp complex cepstrum-based decomposition for asynchronous glottal analysis. In *Interspeech*, 2010.

[36] Baris Bozkurt and Dutioit Thierry. Mixed-phase speech modelling and formant estimation, using differential phase spectrums. In *ISCA, Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis*, 2003.

[37] Thomas Drugman. *Adanvances in Glottal Analysis and its Application*. PhD thesis, University of Mons, Docotral School Musics Signal Processing, Belgium, 2011.

[38] A. Oppenheim and R. Schafer. *Discrete-time Signal Processing*. Prentice-Hall, 1989.

[39] Matti Airas. Tkk aparat: An environment for voice inverse filtering and parametrization. *Logopedics Phoniatrics Vocology*, 33:49–64, 2008.

[40] T. Drugman and T. Dutoit. Glottal closure and opening instant detection from speech signals. In *In Proc. Interspeech*, 2009.

[41] S. Saito and F. Itakura. A statistical method for estimation of speech spectral density and formant frequencies. *Electronic Communiction*, 53-A:36–43, 1970.

[42] Chanwoo Kim, Kwang deok Seo, and Wonyong Sung. A robust formant extraction algorithm combining spectral peak picking and root polising. *EURASP, Journal on Applied Signal Processing*, pages 1–16, 2006.

[43] R. Rabiner Lawrence and Schafer Ronald W. *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, Nj, USA, 1978.

[44] Tuomo Raitio. *Voice Source Modelling Techniques for Statistical Parameteric Speech Synthesis*. PhD thesis, School of Electrical Engineering, Department of Signal Processing and Acoustics, Aalto University, 2015.

[45] D. Y. Wong, J. D. Markel, and J. A. H. Gray. Least squares glottal inverse filtering from the acoustic speech waveform. *IEEE Transcations on Acoustics, Speech, and Signal Processing*, 27:350–355, 1979.

[46] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphons. *Speech Communication*, 9:453–467, 1990.

[47] N. D. Gaubitch and P. A. Naylor. Spatiotemporal averaging method for enhancement of reverberant speech. In *International Conference on Digital Signal Processing (DSP)*, 2007.

[48] Y. Stylianou. Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Transcations on Acoustics, Speech, Audio Processing*, 9:21–29, 2001.

[49] T. Ananthapadmanabha and B. Yegnanarayana. Epoch extraction from linear prediction residual for identification of closed glottis interval. *IEEE Transactions on Acoustics Speech and Signal Processing*, 27:309–319, 1979.

[50] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes. Estimation of glottal closure instants in voiced speech using the dypsa algorithm. *IEEE Transactions on Speech and Audio Processing*, 15:34–43, 2007.

[51] B. Yegnanarayana and R. Smits. A robust method for determining instants of major excitations in voiced speech. In *IEEE Internation Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1995.

[52] K. S. R. Murty and B. Yegnanarayana. Epoch extraction from speech signals. *IEEE Transactions on Speech and Audio Processing,*, 16:1602–1613, 2008.

[53] M. R. P.; Thomas, J.; Gudnason, and P. A. Naylor. Estimation of glottal closure and opening instants in voiced speech using yaga algorithm. *IEEE transactions on Audio, Speech and Language Processing,*, 20:82–91, 2012.

[54] T.; Drugman, M.; Thomas, J.; Gudnason, P.; Naylor, and T. Dutoit. Detection of glottal closure instants from speech signals: A quantitative review. *IEEE Transcations on Audio, Speech, and Language Processing,*, 20:994–1006, 2012.

[55] R. Timcke, Von Leden H., and Moore P. Laryngeal vibrations: measurements of the glottic wave. *Archive. Otolaryngol*, pages 1–19, 1958.

[56] R. Monsen and A. Engebretson. Study of variants in the male and female glottal wave. *Journal of the Acoustical Society of America*, 62:981–993, 1977.

[57] Paavo Alku. Glottal inverse filtering analysis of human voice production- a review of estimation and parameterization methods of the glottal excitation and their applications. *Sadhana*, 36:623–650, 2011.

[58] I. Titze and J. Sundberg. Vocal intensity in speakers and singers. *Journal of the Acoustical Society of America*, 91(5):2936–2946, 1992.

[59] D. G. Childers and C. K. Lee. Vocal quality factors: Analysis, synthesis, and perception. *Journal of the Acoustical Society of America*, 90:2394–2410, 1991.

[60] aand Strike H. Alku, P. and E. Vilkman. Parabolic spectral parameter - a new method for quantification of the glottal flow. *Speech Communication*, 22:67–97, 1997.

[61] P. Howell and M. Williams. The contribution of the excitatory source to the perception of neutral vowels in stuttered speech. *Journal of the Acoustical Society of America*, 84(1):80–89, 1988.

[62] P. Howell and M. Williams. Acoustic analysis and perception of vowels in children's and teenagers's stuttered speech,. *Journal of the Acoustical Society of America*, 91:1697–1706, 1992.

[63] Tuomo Raitio. *Voice Source Modelling techniques for Statistical Parameteric Speech Synthesis*. PhD thesis, Department of Signal Processing and Acoustic, Aalto University, 2015.

[64] T. Yoshimura, K. Tokuda, .T. Masuko, T. Kobayashi, and T. Kitamura. Simultaneous modeling of spectrum, pitch, and duration in hmm-based speech synthesis. In *IN Processdings of Eurospeech*, pages 2259–2262, 1999.

[65] H. Pulakka M. Vainio T. Raitio, A. Suni and P. Alku. Utilizing glottal source pulse library for generating improved excitation signal for hmm-based speech synthesis. In *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2011.

[66] B. Moore and R. Glasberg. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America*, 74:750–753, 1983.

[67] B Weinberg and S. Bennet. A comparison of fundamental frequency characteristics measured on a wave-by-wave and averaging basis. *Journal of Speech and Hearing Research*, 15:351–355, 1972.

[68] N. Sisty and B Weinberg. Vowel formant frequency characteristics of esophageal speech. *Journal of Speech and Hearing Research*, 15:439–448, 1972.

[69] S. Bennet and Weinberg B. Acceptability ratings of normal, esophageal and artificial larynx speech. *Journal of Speech and Hearing Research*, 38:608–615, 1973.

[70] Weinberg B. Feth L. Smith, B. and Y. Horri. Vocal roughness and jitter characteristics of vowels produced by esophageal speakers. *Journal of Speech and Hearing Research*, 21:240–249, 1978.

[71] Horri Y. Weinberg, B. and B. Smith. Long time spectral and intensity characteristics of esophageal speech. *Journal of Acoustic society of America*, 67:1781–1784, 1980.

[72] B Weinberg. Speech after laryngectomy: An overview of acoustic and temporal characteristics of esophageal speech. *Electroacoustic Analysis and Enhancement of Alaryngeal Speech*, pages 5–48, 1982.

[73] B Weinberg. Acoustical properties of esophageal and tracheoesopahgeal speech. Technical report, Laryngectomee Rehabilitation. College Hill Press, San Diego, 1986.

[74] Fisher H. Bolm E. Robbins, J. and M. Singer. A comparative acoustic study of normal, esophageal and tracheoesopahgeal speech production. *Journal of Speech and Hearing Research*, 49:202–210, 1984.

[75] L. Nord and B Hammarberg. Analysis of laryngectomee speech- a progress report. In *In Euorpean Conference on Speech Communication and Technology*, 1989.

[76] M. Trudea and Y. Qi. Acoustical characteristics of female tracheoesopahgeal speech. *Journal of Speech and Hearing Dis*, 55:244–250, 1990.

[77] Q. Yingyong, W. Bernd, and B. Ning. Enhancement of female esophageal and tracheoesophageal speech. *Acoustical Society of America*, 98(5, Pt1):2461–2465, 1995.

[78] Yingyong Qi. Replacing tracheoesophageal voicing source using lpc synthesis. *Acoustical Society of America*, 5:1228–1235, 1995.

[79] Y. Qi and Weinberg B. Characteristics of voicing source waveforms produced by esophageal and tracheoesophageal speakers. *Journal of Speech and Hearing Research*, 38, 1995.

[80] R. Sirichokswad, P. Boonpramuk, N. Kasemkosin, P. Chanyagorn, W. Charoensuk, and H. H. Szu. Improvement of esophageal speech using lpc and lf model. *Internation Conf. on Biomedical and Pharamaceutical Engineering 2006*, pages 405–408, 2006.

[81] R.G. Tull and J.C. Rutledge. Linear predictive synthesis of vowels for pitch enhancement of female geriatric esophageal speech. In *Engineering in Medicine and Biology Society, 1993. Proceedings of the 15th Annual International Conference of the IEEE*, pages 1359–1360, 1993.

[82] M. Kenji and H. Noriyo. Enhancement of esophageal speech using formant synthesis. *Acoustics, Speech and Signal Processing, International conf.*, pages 81–85, 1999.

[83] R. H. Ali and S. B. Jebara. Esophageal speech enhancement using excitation source synthesis and formant structure modification. *SITIS*, pages 615–624, 2006.

[84] A. Loscos and J. Bonada. Esophageal voice enhancement by modeling radiated pulses in frequency domain. *Audio Engineering Society*, 2006.

[85] H. Doi, K.. Nakamura, T. Toda, H. Saruwatari, and K.; Shikano. Statistical approach to enhancing esophageal speech based on gaussian mixture models. *Acoustics Speech and Signal Processing(ICASSP), 2010 IEEE International Conference*, pages 4250–4253, 2010.

[86] P. Sabayjai, P. Boonpranuk, P. Kayasith, and C. Wutiwiwatchai. A study of recognition rate improvement for thai esophageal speech by using feature conversion. In *Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 2009. ECTI-CON 2009. 6th International Conference on*, volume 02 of *1014-1017*, 2009. doi: 10.1109/ECTICON.2009.5137217.

[87] R. A. Prosek and L. L. Vreeland. The intelligibility of time domain edited esophageal speech. *American Speech Language Hearing Association*, 44:525–534, 2001.

[88] Wen-Rong Wu and Po-Cheng Chen. Subband kalman filtering for speech enhancement. *Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on*, 45(8):1072–1083, 1998. doi: 10.1109/82.718814.

[89] Wen-Rong Wu, Po-Cheng Chen, Hwai-Tsu Chang, and Chun-Hung Kuo. Frame-based subband kalman filtering for speech enhancement. In *Signal Processing Proceedings, 1998. ICSP' 98. International Conference*, volume 1, pages 682–685, 1998.

[90] E. Grivel M. Gabrea and M. Najim. A single microphone kalman filter- based noise cancellor. *IEEE Signal processing Letters*, 6(3):55–57, 1999.

[91] Wen-Rong Wu and Po-Cheng Chen. Frame-based sub-band kalman filtering for speech enhancement. *Acoustic Society of America*, 113, 2003.

[92] D. Weixiu and P. Driessen. Speech enhancement based on kalman filtering and em algorithm. *IEEE Pacific Rim Conf. on Communication, Computers and Signal Processing, 1991*, pages 142–145, 1991.

[93] B. Koo J. D. Gibson and S.D Gray. Filtering of colored noise for speech enhancement and coding. *IEEE Trans. , Signal Processing*, 39(8):1732–1742, 1991.

[94] Stephen So and Kuldip K. Paliwal. Modulation-domain kalman filtering for single-channel speech enhancement. *Speech Commun.*, 53(6):818–829, July 2011. ISSN 0167-6393. doi: 10.1016/j.specom.2011.02.001.

[95] Stephen So and Kuldip .K. Paliwal. Supressing the influence of additive noise on the kalman gain for low residual noise speech enhancement. *Elsever, Speech communication 53*, 53:355–378, 2011.

[96] Stephen So and Kuldip .K. Paliwal. Fast converging iterative kalman filtering for speech enhancement using long and overlapped tapered windows with large side lobe attenuation. In *Interspeech(ISCA)-2010, Makuhari, Chiba, Japan*, pages 1081–1084, 2010.

[97] Leandro Aureliano de Silva and Macelo Basalio Joaquim. Noise reduction in biomedical speech signal processing based on time and frequency kalman filtering combined with spectral subtraction. *Computers & Electrical Engineering*, 34: 154–164, 2008.

[98] B. Garcia, I. Ruiz, J. Vicente, and A. Alonso. Formants measurement for esophageal speech using wavelet with band and resoultion adjustment. *IEEE IS-SPIT*, pages 320–325, 2006.

[99] B. Garcia and A. Mendez. Oesophageal speech enhancement using poles stablization and kalman filtering. *ICASSP*, pages 1597–1600, 2008.

[100] O.R Ibon, B. Garcia, and Z. M. Amaia. New approach for oesophageal speech enhancement. *10th International conference, ISSPA*, 5:225–228, 2010.

[101] A.G. Isasi, B. Garcia, and A. M. Zorrilla. Corrective algorithm for esophageal voice cycle detection. *IEEE*, pages 150–155, 2011.

[102] Charles Pascal Clark. Effective coherent modulation filtering and interpolation of long gaps in acoustic signals. Master's thesis, University of Washington, 2008.

[103] Rizwan Ishaq. Adaptive gain equalizer and modulation frequency domain for noise reduction. Maste thesis, School of Engineering, Blekinge Institute of Technology, 2010.

[104] Pascal Clark Les Atlas and Steven Schimmel. Modulation toolbox version 2.1 for matlab. http://isdl.ee.washington.edu/projects/modulationtoolbox/, University of Washington, September 2010.

[105] Richard E. Turner and Maneesh Shani. Probabilistic amplitude demodulation.

[106] L.E. Atlas and Christiaan Janssen. Coherent modulation spectral filtering for single-channel music source sepration. *IEEE International Conference ICASSP.*, pages 461–464, 2005.

[107] C. P. Clark and L. Atlas. A sum-of-product model for effective coherent modulation filtering. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4485–4488, 2009.

[108] P. Clark and L. E. Atlas. Time-frequency coherent modulation filtering of non-stationary signals. *IEEE transaction on Signal Processing*, 45(57):4323–4332, 2009.

[109] Q. Li and L. Atlas. Coherent modulation filtering for speech. *IEEE, Acoustics Speech and Signal Processing, ICASSP*, pages 4481–4484, 2008.

[110] Rizwan Ishaq, Begona Garcia Zapirain, Muhammad Shahid, and Benny Lovstrom. Subband modulator Kalman filtering for signla channel speech enhancement. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7442–7446, 2013.

[111] Rizwan Ishaq and Begona Garcia Zapirain. Optimal subband kalman filter for normal and oesophageal speech enhancement. *Bio-Medical Materials and Engineering*, 24:3569–3578, 2014.

[112] Eiji Yumoto and Wilbur J. Gould. Harmonics-to-noise ratio as an index of the degree of hoarseness. *Acoustic Society of America*, 71(6):1544–1550, June 1983.

[113] Yin. Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals. *Acoustical Society of America*, 102:537– 543, 1997.

[114] M. Alfredo, P.M. Hector, T. Jorge, and O. Patricia. Analysis and recognition of esophageal speech. *Symposium on Signal Processing and Information Technology*, 5:101–106, 2006.

[115] B. Garcia, J. Vicente, A. Alonso, and E. Loyo. Esophageal voices: glottal flow restoration. *Acoustics, Speech and Signal Processing 2005(ICASSP 05)*, pages 141–144, 2005.

[116] B. Garcia, I. Ruiz, A. Mendez, and M. Mendezona. Objective characterization of oesophageal voice supporting medical diagnosis rehabilitation and monitoring. *Computers in Biology and Medicine, Elsevier*, pages 97–105, 2009.

[117] Ning Bi. *Speech Conversion and Its Application to Alaryngeal Speech Enhancement*. PhD thesis, Department of Speech and Hearing Sciences, The University of Arizona, 1995.

[118] Zadeh L.A. Frequency analysis of variable networks. *Proceedings of the IRE*, 38 (3):291–299, March 1950.

[119] J. R. Deller, J. G. Proakis, and J. H. L. Hansen. *Discrete time processing of speech signals*. Macmillan Publishing Company, New York, 1993.

[120] J. R. D. Jr, J. G. Proakis, and J. H. L. Hansen. *Discrete time processing of speech signals*. 1993.

[121] M. H. Hayes. *Statistical Digital Signal Processing and Modeling*. 1996.

[122] Zenton Goh, Kah-Chye Tan, and T.G. Tan. Postprocessing method for suppressing musical noise generated by spectral subtraction. *Speech and Audio Processing, IEEE Transactions on*, 6(3):287 –292, may 1998.

[123] J.P.F. Glas. A differential fm detector for low-if radios. *IEEE-Vechicular Technology Conference VTC.*, 2:658–662, Sep 1999.

[124] M. S. Vinton and L.E. Atlas. Scalable and progressive audio codec. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, volume 5, pages 3277–3280, 2001. doi: 10.1109/ICASSP.2001.940358.

[125] M. Kenji, H. Noriyo, K. Noriko, and H. Hajime. Enhancement of esophageal speech using formant synthesis. *Acoustic. Sci. and Tech.*, pages 69–76, 2002.

[126] Mattias Dhal Nils Westerlund and Ingvar Claesson. Adaptive gain equalizer for speech enhancement. Research report, Blekinge Institute of Technology, Karlskrona, 2002.

[127] S. Shamma. Encoding sound timbre in the auditory system. *IETE J. Res.*, 49(2): 193–205, 2003.

[128] N. Westerlund, M. Dahl, and I. Claesson. Real-time implementation of an adaptive gain equalizer for speech enhancement purposes. *WSEAS.*, 2003.

[129] N. Westerlund, M. Dahl, and I. Claesson. Speech enhancement using an adaptive gain equalizer. *DSPCS.*, 2003.

[130] L. Atlas, Q. Li, and J. Thompson. Homomorphic modulation spectra. *IEEE International Conference on Accoustics, Speech, and Signal Processing (ICASSP)*, 2:761–764, 2004.

[131] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 1st edition, 2004.

[132] N. Westerlund, M. Dahl, and I. Claesson. Speech enhancement using an adaptive gain equalizer with frequency dependent parameter settings. *VTC04*, 2004.

[133] N. Westerlund, M. Dahl, I. Claesson, B. Sallberg, and H. Akesson. Analog circuit implementation for speech enhancement purposes. *Asilomar Conference on Circuits, Systems and Computers.*, 2004.

[134] M. Dahl, I. Claesson, B. Sallberg, and H. Akesson. A mixed analog -digital hybrid for speech enhancement purposes. *ISCAS.*, 2005.

[135] M. Dahl and B. Sallberg. Speech enhancement implementations in the digital, analog and hybrid domain. *Swedish System on Chip Conference*, 2005.

[136] S. Schimmel and L. E. Atlas. Analysis of signal reconstruction after modulation filtering. *Advanced Signal Processing Algorithms, Architectures, and Implementations*, 5910:163–172, 2005.

[137] N. Westerlund, M. Dahl, and I. Claesson. Speech enhancement for personal communication using an adaptive gain equalizer. *Elsevier Signal Processing.*, 85:1089–1101, 2005.

[138] Corina J. van As-Brooks, Florien J. Koopmans van Beinum, Louis C.W. Pols, and Frans J.M. Hilgers. Acoustic signal typing for evaluation of voice quality in tracheoesophageal speech. *Journal of Voice*, 20(3):355– 368, 2006.

[139] Z.H Ling, Y.J Wu, L. Qin, and R.H Wang. Ustc system for blizzard challenge 2006 an improved hmm-based speech synthesis method. *Blizzard Challenge Workshop*, 2006.

[140] C. Plapous, C. Marro, and P. Scalart. Improved signal-to-noise ratio estimation for speech enhancement. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(6):2098 –2108, nov. 2006.

[141] Benny Sallberg, Nedelko Grbic, and Ingvar Claesson. Implementation aspects of the adaptive gain equalizer, 2006.

[142] Steven Schimmel. Sphsc-503 speech signal processing. Technical report, http://isdl.ee.washington.edu/people/stevenschimmel/sphsc503/files/notes11.pdf, 2006.

[143] S. M. Schimmel, K. R. Fitz, and L.E. Atlas. Frequency reassignment for coherent modulation filtering. *IEEE, Acoustics, Speech and Signal Processing, ICASSP*, 5: 261–264, 2006.

[144] Mireia Farrus, Javier Hernando, and Pascual Ejarque. Jitter and shimmer measurement for spkear recog, 2007.

[145] Zwicker-Eberhard Fastl, Hugo. *Psychoacoustic: Facts and Models*. FastI, Hugo, and Eberhard Zwicker, 2007.

[146] Eliathamby Ambikairajah Phu Ngoc Le. Non-uniform sub-band kalman filtering for speech enhancement. *Internation Conference on Signal Processing and Communication Systems (ICSPCS)*, 2007.

[147] S.M. Schimmel. *Theory of Modulation Frequency Analysis and Modulation Filtering with Applications to Hearing Devices*. PhD thesis, University of Washington, 2007.

[148] L. E. Atlas and S. M. Schimmel. Target talker enhancement in hearing devices. *IEEE, ICASSP*, pages 4201–4204, 2008.

[149] B. King and L. Atlas. Coherent modulation comb filtering for enhancing speech in wind noise. *International Workshop on Acoustice Echo and Noise Control*, Sep 2008.

[150] S.M. Schimmel and L.E. Atlas. Target talker enhancement in hearing devices. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 4201 –4204, 31 2008-april 4 2008.

[151] Stephen So and Kuldip .K. Paliwal. A long state vector kalman filter for speech enhancement. In *Proceedings of the 9th Annual Conference of the International Speech Communication Association*, pages 391–394, 2008.

[152] Y. Uemura, Yu. Takahashi, H. Saruwatari, K. Shikano, and K. Kondo. Musical noise generation analysis for noise reduction methods based on spectral subtraction and mmse stsa estimation. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 4433 –4436, april 2009.

[153] Marzena Miesikowska, Leszek Radziszwski, Stanisaw Bien, and Slawomir Oka. Acoustic analysis of esophageal speech internoise 2010. volume 8, pages 6647 – 6655, Lisbon, Portugal, 2010. Acoustic analysis;Esophageal speech;Formant frequency;Formant's loop;Fundamental frequencies;Spectral parameters;Speech signals;Time-based analysis;.

[154] Kuldip Paliwal, Kamil Wójcicki, and Belinda Schwerin. Single-channel speech enhancement using spectral subtraction in the short-time modulation domain. *Speech Commun.*, 52(5):450–475, May 2010. ISSN 0167-6393.

[155] G. Sell and M. Slaney. Solving demodulation as an optimization problem. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(8):2051 –2066, nov. 2010.

[156] Tuomo Raitio, Antti Sunni, Hannu Pulakka, Martti Vainio, and Paavo Alku. Comparison of formant enhancement methods for hmm-based speech synthesis. pages 334–339, September 2010.

[157] B. Huttner, M. Dollinger, G. Luegmair, U. Eysholdt, A. Zeithe, and E. Gurlek. Parameter optimzation for a time depedent multi mass model for the pharyngoesophageal segment. *7th Int. workshop on Models and analysis of vocal emission for biomedical application*, pages 49–52, 2011.

[158] M. O. John and B. Garcia. Quantifiying paramters of a source filter model for oesophageal speech. *IEEE*, pages 532–53, 2011.

[159] Muhammad Shahid, Rizwan Ishaq, Benny Sallberg, Nedelko Grbic, Benny Lovstrom, and Ingvar Claesson. Modulation domain adaptive gain equalizer for speech enhancement. In *Signal and Image Processing Application 2011, by IASTED*, 2011.

[160] A. Alwan. Voicesauce: A program for voice analysis @ONLINE, February 2012. URL http://www.ee.ucla.edu/~spapl/voicesauce/.

[161] R. Ishaq, M. Shahid, B. Lovstrom, B. G. Zapirain, and I. Claesson. Modulation frequency domain adaptive gain equlizer using convex optimization. *6th International Conference on Signal Processing and Communication Systems- 2012*, 2012.

[162] Rizwan Ishaq and Begona Garcia Zapirain. Adaptive gain equalizer for improvement of esophageal speech. *IEEE International Symoposium on Signal Processing and Information Technology (ISSPIT 2012)*, 2012.

[163] Bingyin Xia Na Li Yan Liang, Changchun Bao and Ruwei. An lp spectrum modification method for noisy speech based on linear extrapolation. In *TSP*, pages 450–453, 2012.

[164] B. Tang, A. Shen, G. Pottie, and A. Alwan. Spectral analysis of subband filtered signals. In *Acoustics, Speech and Signal Proessing, 1995, ICASSP-95., 1995 International Conference on*, volume 2, pages 1324–1327, May 1995.

[165] T. Sardjono, R. Hidayati, N. Purnami, A. Noortjaha, G. Verkerke, and M. Purnomo. A priliminary results of voice spectrum analays from laryngectomised patients with and without electrolarynx: A case study in indonesian laryngectomised patients. *Dept. Biomedical Engieering University.*

[166] Teresa Cervera, Jose L. Miralles, and Julio Gonzalez-Alvarez. Acoustical analysis of spanish vowels produced by laryngectomized subjects. *Journal of Speech, Language, and Hearing Research*, 44:988–996, 2001.

[167] Hamid Reza Sharifzadeh, Ian V Mcloughlin, and Ahmadi Farzaneh. Reconstruction of normal sounding speech for laryngectomy patients through a modified celp codec. *IEEE Transaction on Biomedical Engineering*, 57:2448–2458, 2010.

[168] K. Ferrat and M. Guerti. A study of sounds produced by algerian esophageal speakers. *African Health Sciences*, 12:452–458, 2012.

[169] Ian V. Mcloughlin, Hamid Reza Sharifzadeh, Su Lim Tan, Jingjie Li, and Yan Song. Reconstruction of phonated speech from whispers using fromant-derived plausible pitch modulation. *ACM Transcations on Accessible Computing (TACCESS)*, 6:12:1–12:21, 2015.

[170] Peter J. Murphy. Perturbation-free measurement of the harmonics-to-noise ratio in voice signals using pitch synchronous harmonic analysis. *The Journal of the Acoustical Society of America*, 105:2866–2881, 1999.

[171] M. Vainio A. Suni, T. Raitio and P. Alku. The glottalhmm entery for blizzard challenge 2011: Utilizing source unit selection in hmm-based speech synthesis for improved excitation generation. In *in Blizzard Challenge 2011, Workshop, Florence, Italy*, 2011.

[172] Juin-Hwey Chen and A. Gersho. Adaptive postfiltering for quality enhancement of coded speech. *Speech and Audio Processing, IEEE Transactions on*, 3:59–71, 1995.

[173] URL http://www.audio-technica.com/cms/wired_mics/a0933a662b5ed0e2/.

[174] URL http://www.nch.com.au/wavepad/.

[175] Itu-t recommendation p.800 methods for subjective determination of transmission quality.

[176] Gunnar. Fant, Johan. Liljencrants, and Qi-guang Lin. A four-parameter model of glottal flow. Technical report, Dept. for Speech, Music and Hearing, 1985.

[177] Rizwan Ishaq and Begona Garcia Zapirain. Adaptive gain equalizer for improvement of esophageal speech. In *IEEE International Symposium on Signal Processing and Information Technology*, 2012.

[178] Rizwan Ishaq, Muhammad Shahid, Benny Lovstrom, Begona Garcia Zapirain, and Ingvar Claesson. Modulation frequency domain adaptive gain equalizer using convex optimization. In *Signal Processing and Communication Systems (ICSPCS), 2012 6th International Conference on*, pages 1–5, 2012.

[179] Begona Garcia Zapirain Rizwan Ishaq. Esophageal speech enhancement using modified voicing source. In *Signal Processing and Information Technology(ISSPIT), 2013 IEEE International Symposium on*, pages 210–214, 2013.

[180] Rizwan Ishaq, Dhanananjaya Gowda, Paavo Alku, and Begonya Garcia Zapirain. Vowel enhancement in early stage spanish esophageal speech using natural glottal flow pulse and vocal tract frequency warping. In *Proceeding of the 6th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT 2015), Dresden, Germany*, September 2015. URL http://www.slpat.org/slpat2015/papers/ishaq-gowda-alku-zapirain.pdf.