8-25-2016

Hypernasal Speech Analysis via Emperical Mode Decomposition and the Teager-Kasiser Energy Operator

Christopher De La Cruz

Christopher F. De La Cruz

Candidate

Electrical and Computer Engineering

Department

This thesis is approved, and it is acceptable in quality and form for publication: *Approved by the Thesis Committee:*

Dr. Balu Santhanam, Chair

Dr. Amy Neel, Member

Dr. Ramiro Jordan, Member

Hypernasal Speech Analysis via Empirical Mode Decomposition and the Teager-Kaiser Energy Operator

by

Christopher F. De La Cruz

B.S., Electrical Engineering, University of New Mexico, 2013

THESIS

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science
Electrical Engineering

The University of New Mexico

Albuquerque, New Mexico

July, 2016

Dedication

To my family, especially my mother Dolores De La Cruz, for removing life's pressures while I pursued my ambitions.

Also to the memory of my late father Albert De La Cruz, who always knew I would be able to achieve academic success, even if I didn't always know so myself.

Acknowledgments

I would like to thank Dr. Amy Neel from the Speech and Hearing Sciences Department at UNM for her valuable feedback and insights into speech physiology, and to Hillary Jones from the American Cleft Palate-Craniofacial Association for re-establishing access to their database. I would mostly like to thank my advisor Dr. Balu Santhanam for taking me on as his graduate student, for always coming up with fresh ideas, and for remaining patient throughout the research and thesis process.

Hypernasal Speech Analysis via Empirical Mode Decomposition and the Teager-Kaiser Energy Operator

by

Christopher F. De La Cruz

B.S., Electrical Engineering, University of New Mexico, 2013

M.S., Electrical Engineering, University of New Mexico, 2016

Abstract

In the area of speech science, one particular problem of importance has been to develop a clear method for detecting hypernasality in speech. For speech pathologists, hypernsality is a critical diagnostic used for judging the severity of velopharyngeal (nasal cavity/mouth separation) inadequacy in children with a cleft lip or cleft palate condition. For physicians and particularly neurologists, these same velopharyngeal inadequacies are believed to be linked to nervous system disorders such as Alzheimer's disease and particularly Parkinson's disease. One can therefore envision the need to not only find a reliable method for detecting hypernasality, but to also quantify the level (severity) of hypernasality as well.

An integral component in the study of speech is the analysis of speech formants, i.e., vocal tract resonances. Traditional acoustical analysis methods of using a linear source model follow the premise that differences between normal and hypernasal speech can be distinguished by shifts or power changes in the formant frequencies and/or the widening (or narrowing) of the formant bandwidths. Such a premise,

however, has not been validated with consistency. Part of the reason is that traditional acoustical analysis methods such as one-third octave band, LPC (Linear Predictive Coding), and cepstral analysis are ill-equipped to deal with the nonlinear, non-stationary, and wideband characteristics of normal and nasal speech signals. Relatively newer DSP methods that employ group delay or energy separation overcome some of these problems, but have their own issues such as possible mode mixing, noise, and the aforementioned wideband problem. However, initial investigations into energy separation methods show promise as long as these issues can be resolved.

This thesis evaluates the success of a novel acoustical energy approach which deals with the mode mixing and wideband problems where: (1) a DSP sifting algorithm known as the EMD (Empirical Mode Decomposition) is first implemented to decompose the voice signal into a number of IMFs (Intrinsic Mode Functions). (2) Energy analysis is performed on each IMF via the Teager-Kaiser Energy Operator. The proposed EMD energy approach is applied to voice samples taken from the American CLP Craniofacial database and is shown to produce a clear delineation between normal and nasal samples and between different levels of hypernasality.

Contents

List of Figures

List of Tables

Glossary

TKEO	Teager-Kaiser Energy Operator.
IF	Instantaneous Frequency
ESA	Energy Separation Algorithm
CP	cleft palate
CLP	cleft lip and palate.
ACP-CA	*American Cleft Palate-Craniofacial Assoc.*
EMD	Empirical Mode Decomposition.
IMF	Intrinsic Mode Function.
LPC	Linear Predictive Coding
SHS	Speech and Hearing Sciences
IPA	International Phonetic Alphabet
ARPA	Advanced Research Projects Agency
DSP	Digital Signal Processing
FFT	Fast Fourier Transform
DTFT	Discrete Time Fourier Transform

Chapter 1

Introduction

1.1 Overview

A reliable method for acoustically detecting hypernasality in speech has proven to be a tough proposition to develop. Many methods have been tried, often with inconsistent results or general disagreement across different methods. Traditional methods, such as the one-third octave band approach or LPC, assume a narrowband formant model, and are therefore ill-equipped to deal with the wideband nature of normal and nasal speech. Furthermore, the non-linear and non-stationary characteristics of speech cause even more problems for these methods, particularly for LPC. Other methods, such as the ESA or the group delay function, rely on successful multi-band filtering (Chebyshev or Gabor bandpass filter bank) which is also found wanting, since nasal formants are spectrally close to each other and have considerably wide bandwidths. Newer energy-analysis methods with the Teager-Kaiser operator, which are ideal for wideband, non-linear, and non-stationary signals, also fall short because of the same band separation issues.

In this work, we apply a relatively new approach to band separation so that

the Teager-Kaiser energy operator can be applied successfully. The band separation is performed with an algorithm known as EMD (Empirical Mode Decomposition), which is essentially an adaptive sifting algorithm. The EMD is successful where other decompositions are not because it is not based on a set of predetermined analytic functions. Instead, the basis functions for the algorithm are derived adaptively from the data signal itself.

In this thesis, we will examine how some of the traditional speech analysis methods are employed, and why they are not ideal for detecting hypernasality. We will then analyze speech samples from the ACP-CA [28] database using the joint EMD/TKEO method to show that this newer approach is not only more sound for nasal speech analysis, but is also able to produce a much clearer delineation between normal and hypernasal speech.

Outside of evaluating velopharyngeal inadequacy and diagnosing Parkinson's disease, there are other reasons to develop an acoustic method for measuring hypernasality in speech. For the field clinician or therapist, such a tool would provide an inexpensive means of diagnosing and treating nasal-related speech impediments. For instance, such a method could be used to provide real-time biofeedback for a patient training to reduce hypernasality, or for helping a person adapt to speaking a second language.

1.2 Thesis Layout

We begin by introducing terminology and background information for understanding the basics of speech production. Because of its importance, the mechanisms regarding the production of speech formants are covered in detail. Terminology for digital speech processing is covered in a similar manner.

We then look at some traditional, if not entirely successful methods of analyzing

speech signals. The most widely used methods, one-third octave band and LPC are examined. This will serve mainly as a backdrop to the discussion of the newer energy-based methods which follow.

The Teager-Kaiser energy operator is introduced and discussed in detail as it is the crux of the current research. Results of analysis, utilizing the operator in the early stages of the research, are shown. The main point here is to demonstrate that the TKEO has marvellous potential as a formant analysis tool, but is not quite viable for use on speech signals in their pure (unconditioned) form. Voice samples taken from the ACP-CA database are analyzed to make the case.

A method for decomposing (conditioning) the raw speech signal before energy analysis is then introduced. This method is known as EMD (Empirical Mode Decomposition). The details of the evolutionary stages of the EMD algorithm are covered so that the use of the variants of the EMD algorithm can be appreciated.

Finally, the TKEO is applied to the individual sub-components (IMFs) of the EMD and an ensemble-type of energy analysis is performed. The analysis basically defines a set of energy metrics that are derived from the energy contributions of the individual IMFs at pre-determined high and low frequency bands. Voice samples from this section are taken exclusively from the ACP-CA database since the subjects are actual cleft-palate patients. The results of this dual EMD/TKEO approach and recommendations for future work, close the thesis

Chapter 2

Speech Basics

2.1 Anatomy and Differences between Normal and Nasal Speech

As can be seen in Figure 2.1, there are many organs that make up the speech-producing mechanism. For the purposes of studying hypernasality and formant formations, the organs of interest are the velum (soft palate) and the tongue.

In general, human speech is produced by expelling air from the lungs through the vocal folds which vibrate in connection with variations of air pressure in the glottis. This interaction produces quasi-periodic pulses which then pass through the vocal tract. For normal speech, the pulses then travel mainly through the oral cavity and out through the mouth, with only a small amount passing through the nasal cavity. For normal speakers, air flow through the nasal cavity is prevented by closure of the velum. For persons with cleft palate or Parkinsons disease, the velum is malformed (for cleft palate [3]), or is malfunctioning (for Parkinsons [4]), and is unable to close-off the nasal cavity, resulting in most(>60%) of the air flow passing into the nasal cavity and out through the nose. In clinical settings, the percentage of air that

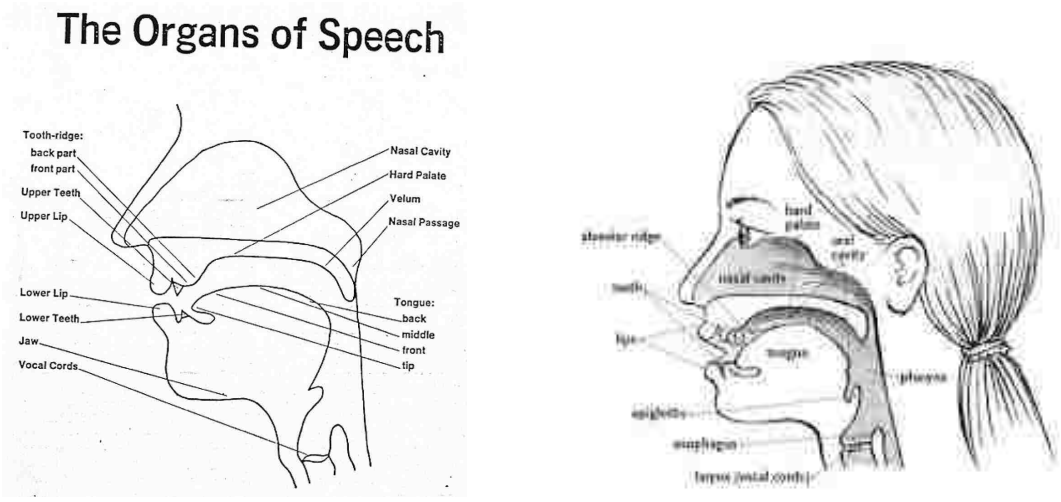


Figure 2.1: Anatomy of the speech-producing mechanism. The soft palate (velum) and the tongue are the vital organs for the production of hypernasal vowels. In the left figure[32], the position of the velum (closed) clearly separates the oral cavity from the nasal cavity resulting in normal speech. In the right figure[33], the velum is slightly lowered (opened), allowing some air to escape through the nasal cavity and out through the nose, resulting in nasal or hypernasal speech.

exits through the nose is measured with a nasometer and produces a percentage score known as *nasalance*. In general, nasalance scores less than 10 indicate normal speech and those above 10 indicate nasal speech. Nasalance scores near 60 and above indicate hypernasal speech.[1]

2.2 Formants

A formant is essentially a resonant frequency formed by the different parts of the speech apparatus, mainly the vocal tract. Different phonemes (units of sound) produce their own unique set of formants which allow scientists to study the different aspects of speech. In the study of normal vs. nasal speech, scientists look for changes in the formant characteristics such as frequency, amplitude, bandwidth, and energy.

[1]As determined from feigned hypernasal voice samples in the SHS lab

If one considers the many ways that the human voice can be utilized (speech, whispering, yelling, singing, etc.), as many as 3 to 6 formants can be present for any particular 'utterance'. For normal (spoken) speech, only the first 3 formants are of interest and it is these that we will focus on.

While formants are present in both vowels and consonants, it is the vowels that are typically used as a speech diagnostic. There are two main reasons for this: First, vowels are well defined sounds and are longer in duration than consonants. Second, the production of vowels utilize parts of the speech apparatus, such as the velum and the nasal cavity, where changes in nasality are more apparent. For instance, the 'ee' sound from the word 'heed' would sound much different between normal and nasal people, as opposed to the 't' sound where the nasal differences would hardly be distinguishable [2].

The first 3 formants for various vowels are shown in Figure 2.2. Note that the formants show up as dark horizontal bands on a spectrogram. For several different languages, the most studied vowels are /i/, /a/, and /u/ (see Figure 2.3 for pronunciations of these and other phonemes). This is not without reason as can be seen from the IPA vowel chart shown in Figure 2.4. The vowel chart is unusually shaped but this is for the purpose of roughly depicting the tongue location for the pronunciation of the various vowels. This is more clearly seen when the chart is overlaid with the vocal tract as shown in Figure 2.5.

As can be seen from Figure 2.5, the geometry of the IPA chart is intended to show the tongue height (positive y-direction) and tongue advancement(negative x-direction) that produce certain vowels. Note that the vowels, /i/, /a/, and /u/ on the corners of the chart represent extreme locations of the tongue, and indeed they are known as corner vowels. In examining our three corner vowels of interest, we note that the 2-dimensional tongue location is close to the roof of the mouth ($\Rightarrow$ high) and is advanced towards the lips ($\Rightarrow$ front) for the vowel /i/, is high and back (away from the lips) for the vowel /u/, and is low (away from the roof of the mouth)

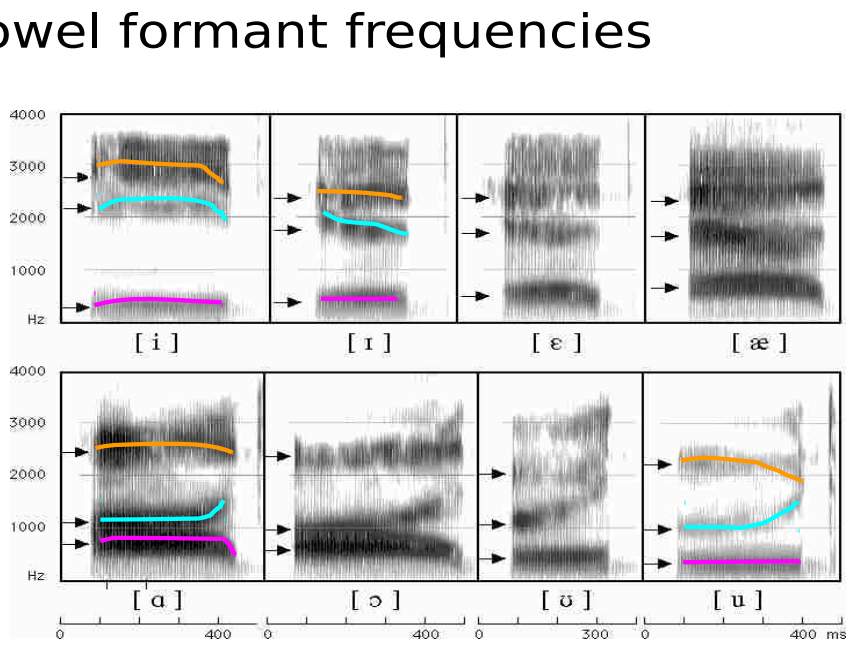


Figure 2.2: Spectrograms showing the first three formants for several vowels [1]. For normal/nasal analysis, the most commonly used vowels are /i/, /a/, and /u/. Reproduced with permission.

and back for the vowel /a/.

It is important to understand the significance of the high/low-front/back tongue locations, as these locations *directly determine the frequencies of the first 2 formants for a particular vowel*. Figure 2.6 shows why this is the case. Note the two high vowels (/i/ and /u/) have approx. the same 1st formant (F1) at about 300 Hz while the two back vowels have 2nd formants (F2) fairly close to each other near 1.1 kHz. A comparison of the three corner vowels leads us to the following general formant rule: *F1 is inversely proportional to tongue height and F2 is proportional to tongue advancement.*

Now that we understand how vowel formants are formed, it remains to determine

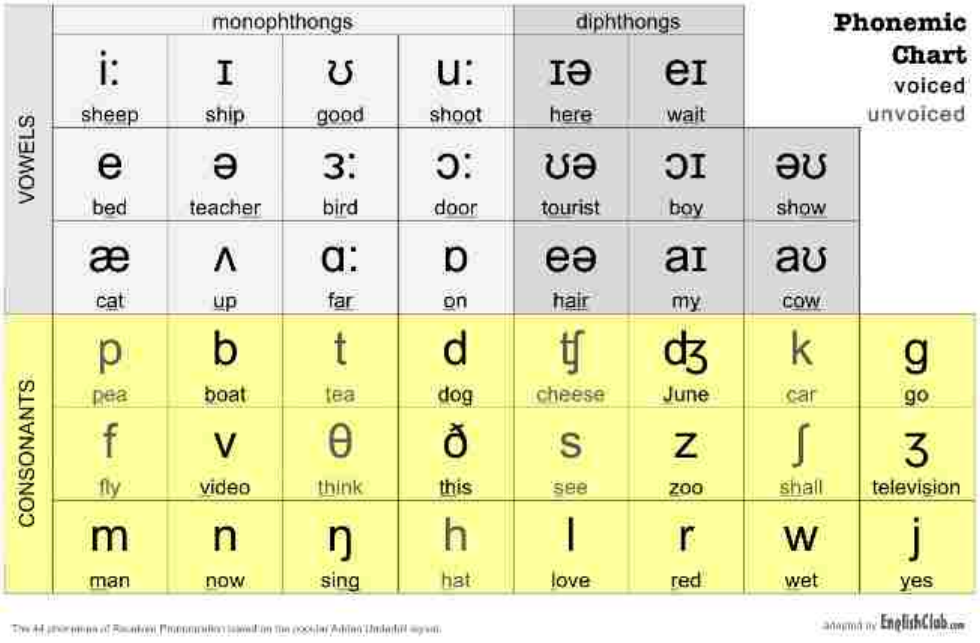


Figure 2.3: Short phoneme chart for common vowels, diphthongs, and consonants [34].

which vowels would be best for the study of hypernasality. Many studies [5, 6, 16, 20, 27] have been done with the vowels /i/, /a/, /e/, and /o/, with the vowel /i/ being subjected to the most study. In fact, /i/ was the only analyzed vowel in [5] and [6]. This seems to be a wise choice, because the frequency separation between the first and second formants is greatest for for this vowel, as can be seen from Figure 2.6 and an expanded list of vowels in Figure 2.7. The initial separation of F1 and F2 is important so that the formant frequency shifts and bandwidth changes in the two formant ranges can be clearly distinguished without spectral overlapping. For this reason, only voice samples of the vowel /i/ will be analyzed in this work.

Returning to Figure 2.7, it should be noted that the formant locations are different between men, women, and children. In general, the formant groups for the men are the lowest, while the formant groups for the children are the highest. This is no doubt due to the normal speaking pitches of the different groups. Also adding to the spread of formant locations, are the inherent differences of the speech apparatuses

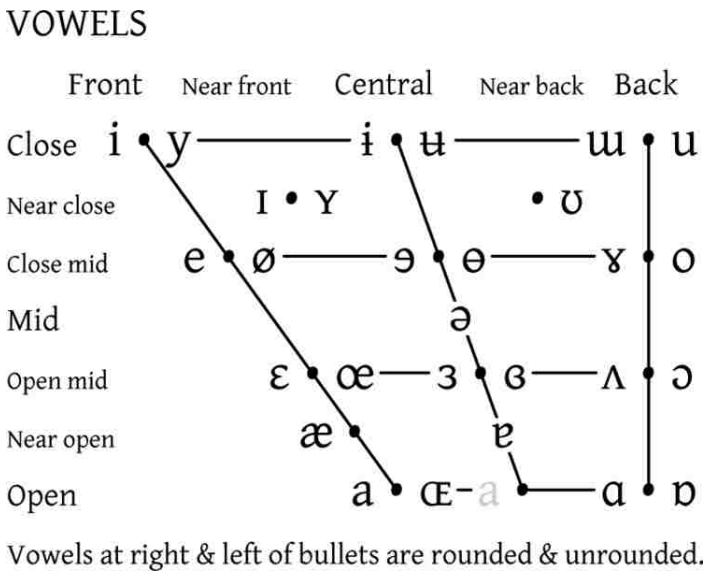


Figure 2.4: IPA vowel chart [35]. The trapezoidal shape is intended to depict extreme ranges of tongue position. The vowels /i/, /a/, and /u/ at the corners of the chart are referred to as 'corner vowels'. Roundness refers to the rounding of the lips during vowel articulation.

Figure 2.5: Overlay of IPA vowel chart and F2 vs. F1 graph with vocal tract [36]. From the chart and the vocal tract we can see why /i/ is termed a 'high/front' vowel. Note that the higher tongue position produces a lower F1 formant and that a forward tongue position produces a higher F2 formant.

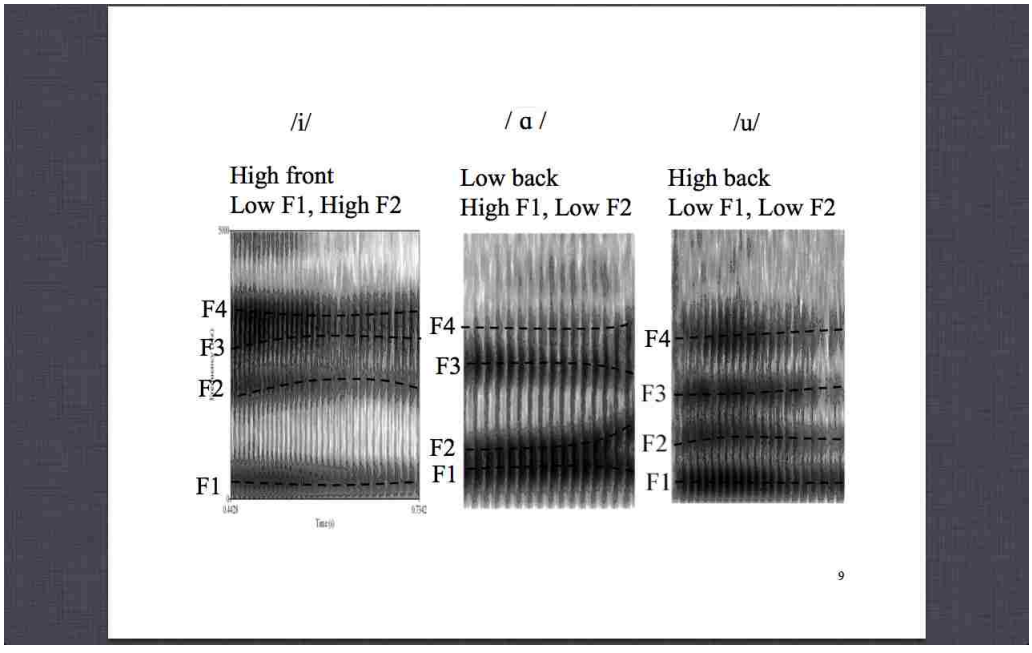


Figure 2.6: Formant comparisons of corner vowels [37]. Note that because of common tongue positioning during vowel articulation, /i/ and /u/ have the same first formant while /a/ and /u/ have approx. the same second formant.

between speakers. In an attempt to bound the ranges for F1 and F2, spectrogram data was collected for several vowels for men and women [2]. This data is shown in

	Men			Women			Children		
Vowel in	F1	F2	F3	F1	F2	F3	F1	F2	F3
beat	270	2300	3000	300	2800	3300	370	3200	3700
bit	400	2000	2550	430	2500	3100	530	2750	3600
bet	530	1850	2500	600	2350	3000	700	2600	3550
bat	660	1700	2400	860	2050	2850	1000	2300	3300
part	730	1100	2450	850	1200	2800	1030	1350	3200
pot	570	850	2400	590	900	2700	680	1050	3200
boot	440	1000	2250	470	1150	2700	560	1400	3300
book	300	850	2250	370	950	2650	430	1150	3250
but	640	1200	2400	760	1400	2800	850	1600	3350
pert	490	1350	1700	500	1650	1950	560	1650	2150

Figure 2.7: Formant locations of corner and additional vowels for men, women, and children [38]. The wide separaton between F1 and F2 for the vowel /i/ ('beat') for all three catagories make it the ideal vowel for formant analysis.

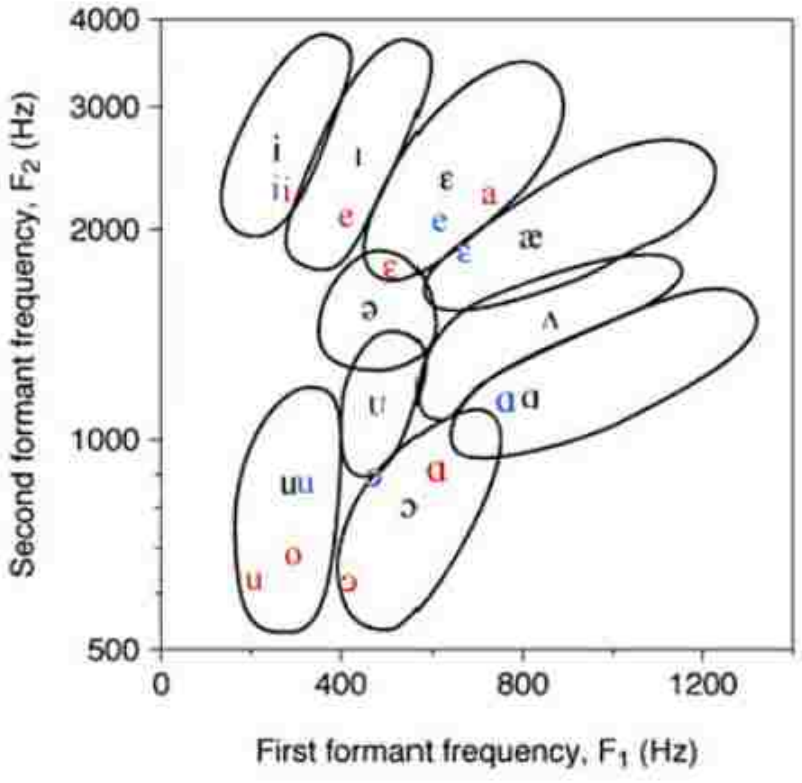


Figure 2.8: F2 vs. F1 plot for several vowels [39]. Data is taken from a wide range of speakers. Values for F1 and F2 vary because of gender, age, and inherent anatomical differences (particularly regarding the vocal track) between subjects. For the vowel /i/, the F1 formant might lie between 300 ± 150 Hz while the F2 formant might lie between 2575 ± 900 Hz.

Figure 2.8. From the figure, we note that a possible range of F1 for the vowel /i/ could be 300 ± 150 Hz while the F2 range could be 2575 ± 900 Hz.

2.3 Anti-Formants and Criteria for Determining Hypernasality

At this point, we have a good idea of where the first two formants of the normally (non-nasal) spoken vowel /i/ are to be found. What about the hypernasal form

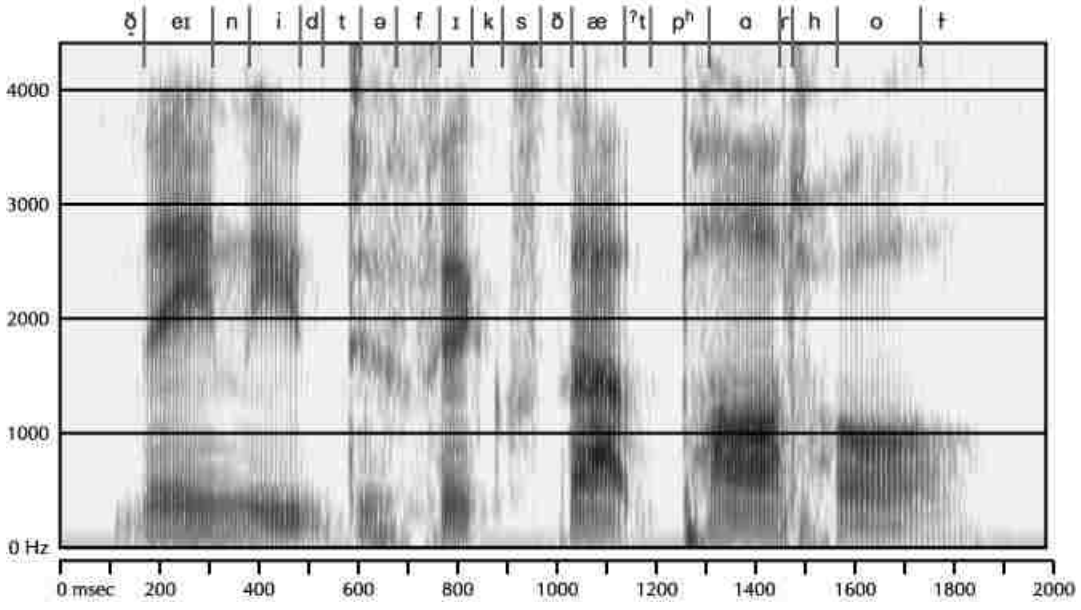


Figure 2.9: Wideband spectrogram for several vowels and consonants [40]. Note that the intensity for the nasal /n/ at 2.6 kHz is significantly lower than the succeeding vowel /i/. To a lesser degree a similar intensity difference can be seen at 3.8 kHz. The lower-intensity of the /n/ indicates the formation of an anti-resonance, or anti-formant, due to the increased interaction between the nasal and oral cavities.

of this vowel? The answer to this is not universally agreed upon in the literature. Before determining a set of criteria for determining hypernasality, we need to explore the counterpart to the formant; the anti-formant.

As the term implies, an anti-formant (anti-resonance) is a weakening of a formant frequency. In our study of hypernasality, the anti-resonance is formed by the increased interaction of the nasal cavity with the oral cavity [5, 6]. In normal speech, this naturally occurs for the nasal consonants (nasals) /n/,/m/, and NX (as in sing in the ARPA chart), where the velum is lowered while the opposite side of the oral cavity is simultaneously closed-off by the lips (/m/) or the tongue (/n/). This oral/nasal interaction is modelled in [2] as a quarter-wavelength resonator, $F_z = c/4I_m$, where I_m is the length from the velum to the closure point (i.e., oral cavity length), and F_z is the frequency related to a newly-formed zero in the vocal tract transfer function.

With the resonator model, an oral cavity length of 3.4 cm should produce a spectral zero, or lower resonance, at approx. 2.6 kHz. This should be kept in mind as we now look at the effects of this interaction in the spectrogram of Figure 2.9.

The spectrogram shows the frequency spectrum for several vowels and consonants. Here were are mainly interested in the nasal consonant /n/ and the succeeding vowel /i/. At approx. 2.6 kHz (vertical scale), the spectral intensity of /n/ is significantly lower than the vowel /i/ following it (and the diphthong /ei/ preceding it). This is consistent with the zero-frequency of the quarter-wavelength resonator. To a lesser degree, the same differences can be seen near 3.8 kHz. We thus confirm that nasals produce lower resonances, or anti-resonances at certain formant frequencies. We expect that this same behaviour will also occur for hypernasal speech, and to a greater degree. From this point forward, we concern ourselves with verifying this last point and with identifying the frequencies an/or frequency ranges where the formants and anti-formants occur for hypernasality.

To this end, we will start by following the one-third octave band criteria from [5] that *the spectral amplitude increases in the region between F1 and F2 (specifically between 630 and 1000 Hz) and decreases in the F2 region (specifically near 2500 Hz).*

Chapter 3

Traditional Methods of Speech Analysis

3.1 One-Third Octave Band Method

One of the earliest methods for evaluating speech hypernasality was the one-third octave spectra analysis approach. Initially developed by Katoaka et al. in the late 1980s [8], this method was widely used until at least 2009. In fact, the results of the hypernasality studies in [5], where the hypernasality criteria for this work were established in Chapter 2, were obtained with the one-third octave band method. The analysis from [6] employed the one-third octave band method as well.

Essentially, the method determines the spectral amplitudes for a pre-determined set of frequency bands at a pre-determined set of center frequencies. The center frequencies are determined by dividing the audible frequency range (10 Hz - 20 kHz) into one-third octave increments using 1 kHz as the initial reference. For instance, a partial set of center frequencies (in Hz) would be 500, 530, 800, 1000, 1250, 1600, 2000. Bandwidths are selected so that the full frequency spectrum is covered without

spectral overlap. A full table of center frequencies and bandwidths can be obtained from [41].

The premise for dividing the audible scale in this manner seems to be a logical one as it is based on the perception of sound by the human ear [2]. Also, by analyzing small number of relatively wide frequency bands as opposed to a large number of narrow frequency bands, the amount of information is compressed.

While these are certainly reasons for embracing the one-third octave band method, we must note that the scale is devised for human *hearing*, but not necessarily human *speech*. The fact that this could be a problem is pointed out in [8] where correlation based on perceptual evaluation was only 50%.

Although the hypernasality criteria established by this method is followed in this work, we believe that the criteria can be further validated and refined by the energy-based methods to follow. We therefore end the discussion on the one-third octave band method, noting that it could be a viable method for hypernasality analysis. However, it cannot delineate between different levels of hypernasality, as opposed to the energy methods that will be investigated.

3.2 LPC Method

3.2.1 LPC Background

LPC (Linear Predictive Coding) has been the predominant tool in speech analysis because of its ability to estimate parameters of speech that can be represented in a discrete-time model. Such a model is shown in Figure 3.1. Due to the cascade (series) nature of the model, a voiced speech utterance can be represented as:

$$S(z) = P(z)G(z)V(z) = E(z)V(z) \tag{3.1}$$

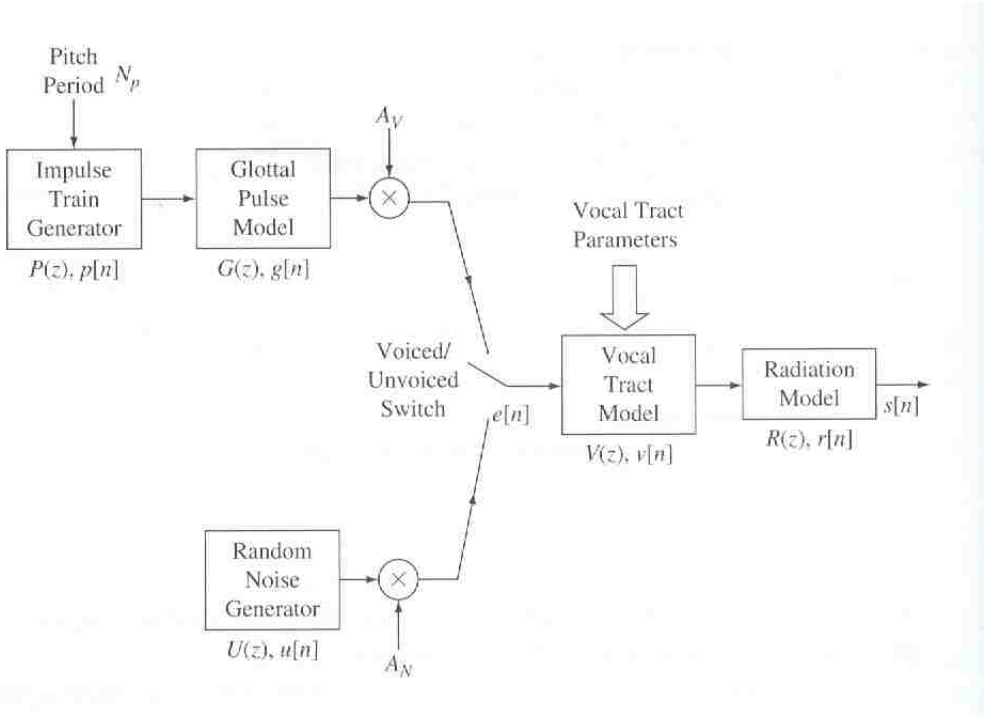


Figure 3.1: Discrete-time model for speech production [2]. Each of the major speech sections can be represented by a spectrum or transfer function in the frequency domain allowing for cascade combination and analysis as a single all-pole system. The all-pole transfer function is suitable for LPC analysis.

In the study of hypernasal speech, we are mainly interested in the transfer function of the vocal tract V(z) which leads to:

$$V(z) = \frac{S(z)}{E(z)} \tag{3.2}$$

Various sources [2, 9] have used a lossless tube model, such as that shown in Figure 3.2, to represent the vocal tract. If the tube lengths (Δx) are all the same, the vocal tract can be represented as a constant delay, all-pole discrete filter represented by

$$V(z) = \frac{1}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}}, \alpha_k = 1, 2, ..., p \tag{3.3}$$

The LPC method works on the principle that a speech sample can be approximated as a *linear combination* of p past speech samples. Linear coefficients are

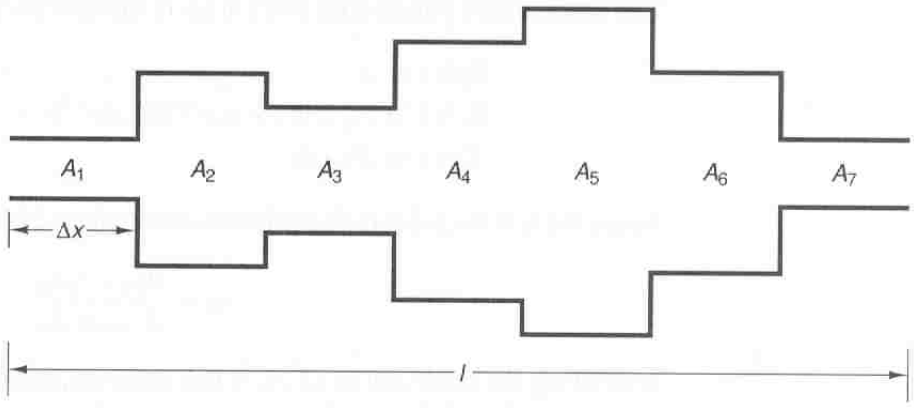


Figure 3.2: Vocal tract model [9] for V(z) from Figure 3.1. The vocal tract can be modelled as a series of lossless tubes with constant length (Δx) and varying tube areas.

approximated minimizing the mean-squared differences between the actual and linear predicted voice samples. In theory, the LPC method will derive the best set of predictor coefficients (α_k) to estimate the time-varying spectral properties (in this case, formants) of a speech signal. While this may be the case for normal speech signals, in the next section we will see that this may not be the case for a wideband nasal speech signal.

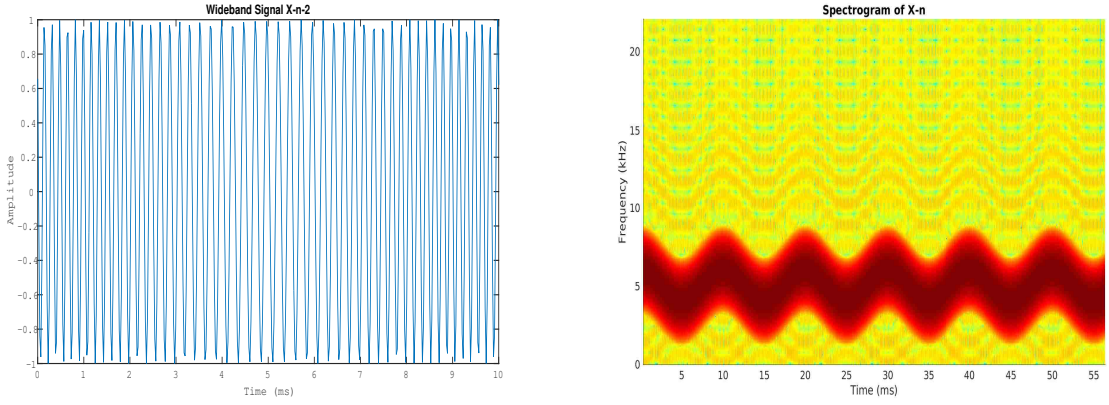
3.2.2 LPC analysis of a wideband signal

In this section, the LPC method is tested on a composed wideband FM signal that is representative of a hypernasal speech signal. The continuous form of the signal is

$$x(t) = \cos \int_{-\infty}^{t} \omega_i(\tau) d\tau, \qquad \text{where} \tag{3.4}$$

$$\omega_i(t) = \omega_c + \omega_m \cos(\omega_f t), \qquad \text{where} \tag{3.5}$$

ω_c, ω_m, and ω_f are the carrier, modulation, and message angular frequencies respectively.



(a) (b)

Figure 3.3: (a) Wideband FM signal for testing LPC. $\Omega_c/\Omega_m = 5$, $\beta = 10$. (b) Spectrogram of wideband test signal in (a). Dominant frequency bands are $\Omega_c \pm k\Omega_m$ (5 kHz $\pm$ k$\times$1 kHz).

The discrete form of the signal is

$$x[n] = \cos \sum_{-\infty}^{n} \Omega_i[m], \qquad \text{where} \tag{3.6}$$

$$\Omega_i[m] = \Omega_c + \Omega_m \cos(\Omega_f[m]), \qquad \text{where} \tag{3.7}$$

Ω_c, Ω_m, and Ω_f are the carrier, modulation, and message angular frequencies respectively.

For the composed signal, $f_c = 5$ kHz, $f_m = 1$ kHz, $f_f = 100$ Hz, and the sampling frequency is 44.1 kHz. The signal is wideband by the criteria $\beta = \Omega_m/\Omega_f = 10 > 1$ [25]. The signal and its spectrogram are shown in Figure 3.3. From the spectrogram we note that the dominant frequency bands are $\Omega_c \pm k\Omega_m$ (5 kHz $\pm$ k$\times$1 kHz).

Figure 3.4 shows a DFT of the composed signal. The DFT gives us a little more information about the frequencies near 5 kHz. In addition to the $\pm$ 1 kHz modulation frequencies centered about the carrier, frequencies of $\pm$ 100 Hz and $\pm$ 200 are present as well. These appear to be multiples of the message frequency Ω_f.

The LPC method was then applied to the composed signal for polynomial orders

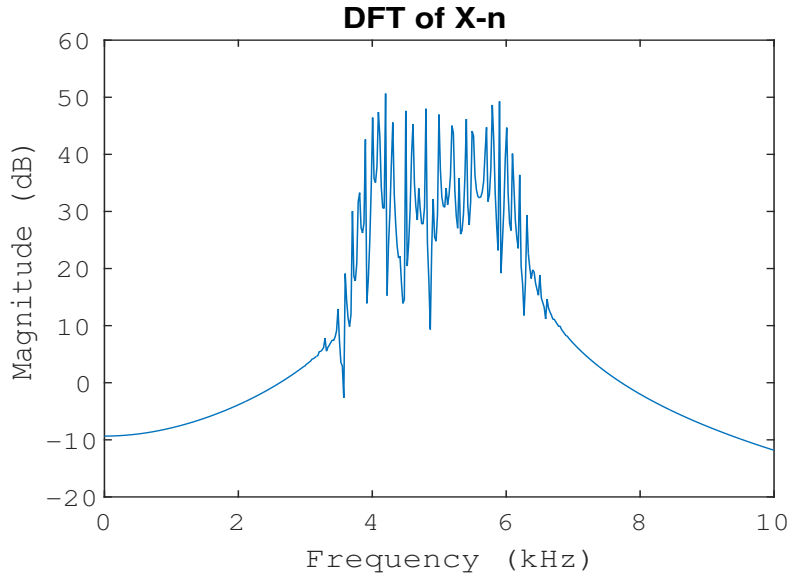
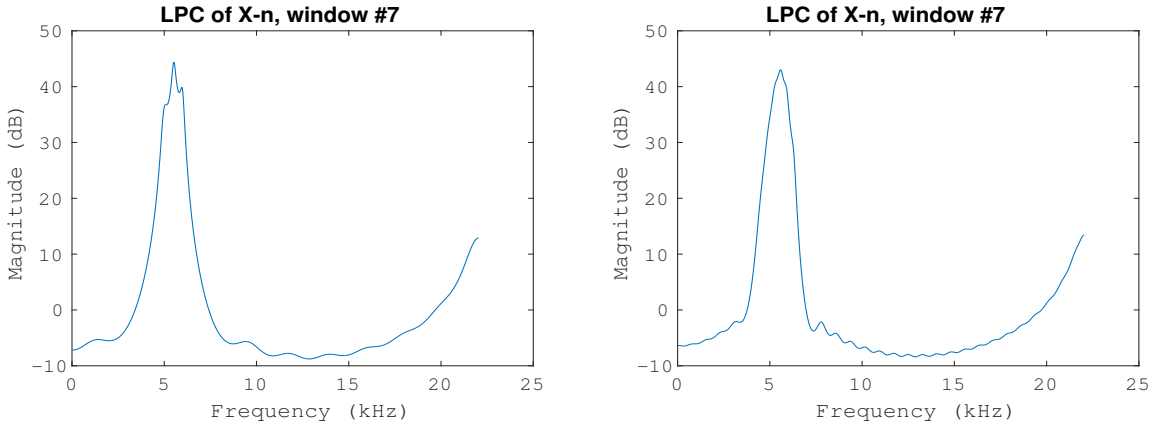


Figure 3.4: DFT of wideband signal $\Omega_c/\Omega_m = 5$, $\beta = 10$. The DFT detects the same frequencies as the LPC plus additional frequencies at $\Omega_c \pm$ 100 - 200 Hz.

of 24 and 60. The order of 24 (for 12 coefficients not including conjugate roots) was chosen because is typically used for formant tracking [27]. The order of 60 was chosen for comparison. In order to evaluate the success of the LPC, we use the criteria from [25] that for wideband signals, dominant frequencies are to be found at $\Omega_c \pm k\Omega_m$, which with the chosen parameters will be ..., 3 kHz, 4 kHz, 5 kHz, 6 kHz, 7 kHz,...

The LPC frequency responses are shown in Figure 3.5. The detected frequencies for each LPC order are tabulated in Table 3.1. In examining the table, there is one disturbing item:

For both LPC orders, the detected frequencies are not in 1 kHz increments from the carrier frequency but instead are in 400 Hz - 500 Hz increments. Clearly, the LPC process is detecting frequencies which are not present in the wideband signal. In a speech analysis scenario, these extra frequencies might be interpreted as extra formants. While the extra frequencies (formants) near f_c would not be detected by a lower-order LPC, it is also possible that legitimate frequencies (formants) farther



(a) (b)

Figure 3.5: LPCs of wideband signal $\Omega_c/\Omega_m = 5$, $\beta = 10$. (a) LPC order = 24. (b) LPC order = 60. Note that both LPCs detect Ω_c in addition to several low-magnitude 'erroneous' frequencies. This is particularly noticeable for LPC-order = 60 at f < 5 kHz (See Table 3.1).

Order	f1	f2	f3	f4	f5	f6	f7	f8
24	—	—	—	—	—	4112	4614	5178
60	709	1492	2291	3713	4161	4497	4840	5218

Table 3.1: Detected frequencies for 24th and 60th order LPCs. Frequencies from single-sided spectrum are reported for brevity. For the composed wideband signal $\Omega_c = 5$ kHz and $\Omega_m = 1$ kHz, detected frequencies should be $\Omega_c \pm k\Omega_m$ (approx. 5 kHz, 4 kHz, 3 kHz,...) but instead are $\Omega_c \pm \frac{1}{2}k\Omega_m$ (approx. 5 kHz, 4.5 kHz, 3 kHz,...).

away from f_c could also go undetected.

In fact this last point is driven home when we look at the range of frequencies from f1 - f4 (not to be confused with F1 and F4 for formant designations) for the 60th order LPC. While this set of frequencies are closer to the expected frequency distribution, they are not easily detectable with a lower-order LPC as can be seen from the table. However, their successful detection with a higher-order LPC comes at the price of detecting other non-existent frequencies!

While this short analysis could be extended to other large frequency deviation scenarios, by itself it serves to demonstrate the danger of using the LPC for analysis

of wideband signals. Here, its failure to produce an unambiguous result for a simple wideband FM problem leads one to explore different methods that can accomplish this objective.

Chapter 4

The Teager-Kaiser Energy Operator (TKEO)

4.1 Background

4.1.1 Teager's early publications and Kaiser's formalization of the energy operator.

"A Phenomenological Model for Vowel Production in the Vocal Track" [11], was published by Teager and Teager in 1983. In this article, the authors made the claim that the speech model of the time was inadequate. The main part of their argument was that the model was based on linear filter theory whereas the actual production of speech was a nonlinear process. In a subsequent publication [12], the authors addressed the nonlinearties of speech in more detail and showed a plot of the "energy" source for the speech signal. No derivation of the energy was given.

Teager's work caught the attention of Kaiser, who in 1990 published an article entitled "On a Simple Algorithm to Calculate the 'energy' of a Signal"[14]. In this

work, he derived the formal algorithm (shown in then next section) for Teager's energy source which he named "Teager's Energy Algoritim". This was very generous on Kaiser's part, considering that he derived the algorithm alone as Teager was never forthcoming on revealing the details supporting his initial energy calculation. Of note here is that Kaiser derived the discrete form of the operator in this publication and only later extended it to the continuous form.

4.1.2 Derivation of operator

While the Teager-Kaiser energy operator (TKEO) has continuous and discrete forms [13], the discrete form is of primary interest here, since we are analyzing digitized speech signals. However, the continuous form is necessary for understanding the physical nature of the operator and will aid in deriving the discrete form. The continuous and discrete forms of the operator are:

$$\Psi(x(t)) = \dot{x}^2 - x(t)\ddot{x}(t) \qquad \text{continuous} \tag{4.1}$$

$$\Psi(x[n]) = x^2[n] - x[n-1]x[n+1] \qquad \text{discrete.} \tag{4.2}$$

In Kaiser's derivation, the energy operator was believed to track the energy behaviour of a harmonic oscillator. An analysis of a simple spring/mass system will show why this is so.

The homogeneous differential equation for a classical spring/mass system is:

$$\frac{d^2x}{dt^2} + \frac{k}{m}x = 0 \tag{4.3}$$

where: x is position, k is the spring constant, m is mass.

The solution of the differential equation is:

$$x(t) = A\cos(\omega t + \phi) \tag{4.4}$$

where: x(t) is position at time t, $\omega = 2\pi f$, ϕ= phase.

The total system energy is the sum of the potential and kinetic energy given by

$$E_{tot} = \frac{1}{2}kx^2 + \frac{1}{2}mv^2. \tag{4.5}$$

Substituting

$$v = \frac{dx}{dt} \text{ and } x = A\cos(\omega t + \phi) \text{ into the above equation yields} \tag{4.6}$$

$$E_{tot} = \frac{1}{2}m\omega^2 A^2 \Rightarrow E_{tot} \equiv A^2\omega^2. \tag{4.7}$$

From this analysis, it is apparent that the energy of a harmonic oscillator is proportional to the oscillating amplitude and frequency.

Restating the continuous form of the TKEO:

$$\Psi(x(t)) = \dot{x}^2(t) - x(t)\ddot{x}(t), \tag{4.8}$$

and substituting x(t) = Acos(ωt + ϕ) from equation 4.4 yields

$$\begin{aligned} \Psi(x(t)) &= (-A\omega\sin(\omega t))^2 - A\cos(\omega t)(-\omega^2 A\cos(\omega t)) \\ &= A^2\omega^2[\sin^2(\omega t) + \cos^2(\omega t)] \\ &= A^2\omega^2. \end{aligned} \tag{4.9}$$

which is the same result as Equation 4.7. It can be surmised that the behaviour of the continuous TKEO is indeed the same as the basic harmonic oscillator. Extending to the discrete case, we rewrite the continuous solution in equation 4.4 in discrete form:

$$x[n] = A\cos(\Omega n + \phi), \tag{4.10}$$

where Ω= 2πf/F, F = sampling frequency.

Since equation 4.10 has 3 unknowns (A, Ω, ϕ), 3 instances of the equation can be used to define a system of linear equations for solution of the unknowns. We define

the instances at an arbitrary point x[n], and at points on either side of x[n], namely at x[n-1] and x[n+1]. This produces the following system of equations:

$$\begin{aligned} x[n] &= A\cos(\Omega[n] + \phi) \qquad (a) \\ x[n-1] &= A\cos(\Omega[n-1] + \phi) \quad (b) \\ x[n+1] &= A\cos(\Omega[n+1] + \phi) \quad (c) \end{aligned} \tag{4.11}$$

Using substitution and the following trigonometric identities:

$$\cos(\alpha+\beta)\cos(\alpha-\beta) = \frac{1}{2}[\cos(2\alpha) + \cos(2\beta)], \tag{4.12}$$

$$\cos(2\alpha) = 2\cos^2(\alpha) - 1 = 1 - 2\sin^2(\alpha), \tag{4.13}$$

yields after multiplying equations 4.11(b) and (c),

$$\begin{aligned} x[n-1]x[n+1] &= A^2\cos^2(\Omega+\phi) - A^2\sin^2(\Omega) \\ &= (x[n])^2 - A^2 sin^2(\Omega). \end{aligned} \tag{4.14}$$

Substituting equation 4.11(a) for x[n] into Equation 4.14 and manipulating we get

$$A^2\sin^2(\Omega) = x^2[n] - x[n-1][n+1], \tag{4.15}$$

which is close to Kaiser's discrete TKEO. In this form, equation 4.15 is exact but yields a unique result only under the condition that Ω is less than $\pi/2$, or that f/F is less than 1/4. With a sampling frequency of 44.1 kHz, this would mean that energies at frequencies above 11 kHz would be unsuitable for analysis.

To get to Kaiser's final form of the operator equation, the left-hand side of equation 4.15 can be simplified by imposing a stronger constraint on Ω conditions. A sampling frequency of 44.1 kHz would mean that energies at frequencies above 5.5 kHz will have at least a 11% error. Assuming we accept this tolerance (as Kaiser did), then

$$A^2\sin^2(\Omega) \equiv A^2\Omega^2, \tag{4.16}$$

and from Equation 4.7 we get

$$A^2\Omega^2 \equiv E_{tot} \equiv \Psi(x(t)) \equiv \Psi(x[n]), \tag{4.17}$$

which in conjunction with Equation 4.15 gives us Kaiser's energy operator equation with the constraint

$$\Psi(x[n]) = x^2[n] - x[n-1]x[n+1], \text{ under the condition } \Omega < \frac{\pi}{4} \tag{4.18}$$

4.1.3 Energy operator characteristics and conditions for non-negativity

As can be seen in Chapter 4.1.2, the energy operator works on a very small (minimal) time scale, i.e., has excellent time-resolution, which makes it ideal for analysis of non-stationary wideband signals, such as human speech. Such was not the case for LPC, as was demonstrated in Chapter 3.

The previous analysis would also seem to suggest that the operator would only be useful for sinusoidal signals but such is not the case. In [16] it was demonstrated that accurate energy signatures could be extracted from exponential, exponentially damped sinusoid, complex (2-dimensional), FM, and AM-FM signals. In fact, the operator is the workhorse for the ESAs (Energy Separation Algorithms) developed in [16] and [17] where the energies of the amplitude and frequency components are isolated. This opens the door to a completely different approach to speech analysis which will not be covered here.

It is important to realize that the operator is intended to model the energy of a *signal source*, not the energy of the *signal* itself. Because of this characteristic, there are certain types of signals where the use of the operator has its limitations. In such cases, the algorithm has been known to produce negative energy values, which would appear to be a physical irregularity. For instance, signals with multiple AM-FM components that differ significantly in amplitude are susceptible to this type of

misinterpretation. Examples where this has occurred [10, 18, 19], would be for a signal like:

$$s(t) = A_1 \cos(\omega_1 t) + A_2 \cos(\omega_2 t), \text{ where } A_1 = 10A_2 \text{ and } \omega_2 = 8\omega_1 \tag{4.19}$$

which is a combination of two sinusoidal signals, which could imply the existence of two different energy sources with significantly different distances. This could account for the significant amount of negative energy as the operator function is geared to compute the energy for a single energy source. This was also seen in [17] where the parameters of $\omega_2 = 1.133\omega_1$ produced a 'beating' frequency resulting in sharp IF reversals from the ESA due to the input of an 'irregular' energy signature.

While this type of error must always be taken into account, it is not an issue for the current work. All of the analyzed voice signals were shown to have negligible negative-energy amplitudes. Furthermore, with the later use of the EMD in Chapter 5, this problem is eliminated, as the main purpose of the EMD is to separate the signal of interest into several 'single-source' subcomponents.

Chapter 5

Early Implementation of the TKEO for Formant Detection

Now that we have a clear understanding of the TKEO, it is time to see how it can be implemented as a formant detector. It should be noted that the operator is not applied directly to a speech signal but rather to resonant subcomponents (frequency bands) of the signal that may or may not be formants. The information contained in the frequency bands depends on what type of voice model is being used and on what type of separator (filter) is applied to the voice signal.

For this chapter, we start from the premise that speech resonances can be modeled as an amplitude and frequency-modulated signal (AM-FM). Work done by Maragos, et al. in [16] demonstrates the validity of this premise. From here the following processes take place: the resonances of the modulated speech signal are separated by a bank of band-pass filters known as Gabor filters. The Gabor filter is optimum for this type of separation and will be explained in detail in the next section. At this point there are two different directions that can be taken:

(1) Each frequency band can be individually processed by an energy separation

algorithm (ESA) which basically demodulates the speech signal into its instantaneous AM and FM components. From these components, the short-time formant frequencies are estimated and are represented in a 'piknogram'. This was done in [16] and [17] with inconclusive results and will not be done here. It should be noted that the ESA utilizes the TKEO.

(2) We can calculate the energy as a function of time for each frequency band with the TKEO to derive an energy spectrum and represent the result as an 'energygram' [15]. The energygram is a 3-dimensional representation of the energy intensity as a function of time and frequency and will serve as a visual aid for determining the formant frequencies. It is our belief that the frequencies with the highest energy levels are indeed formants. This is the method that will now be examined with two speech samples from the ACP-CA database: (1) WOMENRS1 which is a normal signal and (2) WOMENRS6 which is a Level-6 hypernasal signal.

5.1 Gabor Filtering

Before applying the TKEO, the resonances around possible formants are isolated with a bank of Gabor filters. The Gabor filter is a one-dimensional band-pass filter that is cosine-shaped with a Gaussian envelope. In 1946, Gabor demonstrated that this filter design obtained the optimal compromise between localization in the time and frequency domains in compliance with the uncertainty principle [21]. In filter terminology such a filter is described as being compact. Another redeeming quality of the Gabor filter is that the production of large sidelobes is avoided due to its Gaussian shape.

Processing by a bank of filters means that the speech signal is filtered at many different center frequencies, which in our case translates to formant candidate frequencies. For instance, if we want to search for formant candidates in the range of

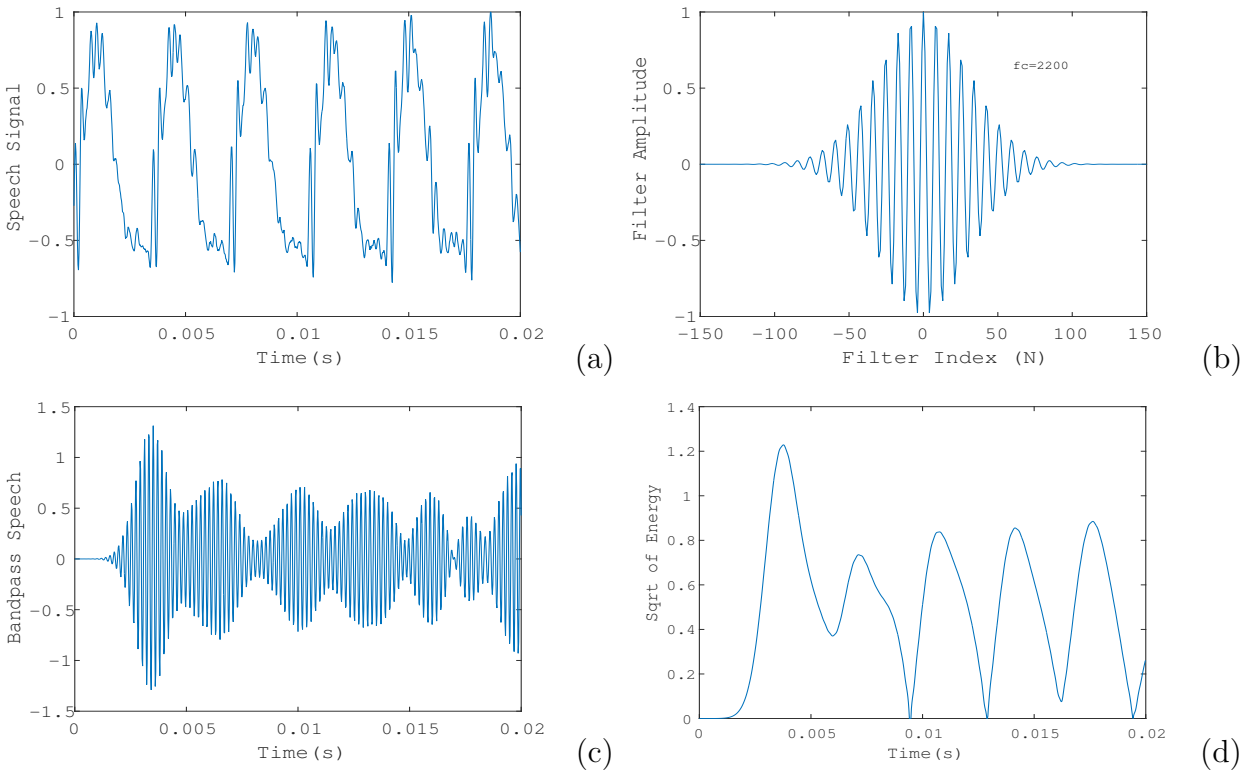


Figure 5.1: Processing steps of single-filterbank energy output for signal WOMENRS1. (a) time-truncated speech signal. (b) time-domain Gabor filter impulse response at bandpass f_c of 2,200 Hz and BW of 400 Hz. (c) Gabor-filtered speech signal. (d) square root of energy. The full Gabor filter bank (250 filters) sweeps the range 200 Hz - 5.2 kHz in 20 Hz increments. All energy components are later combined to form an 'energygram'.

200 Hz to 4.2 kHz, we could set up a bank of 80 Gabor filters to cover this range at 50 Hz increments. For the following analysis, a bank of 250 Gabor filters is used to cover the range 200 Hz to 5.2 kHz in 20 Hz increments.

For signal processing, along with a range of center frequencies (f_c), a filter bandwidth (BW) must be specified. Experiments in [16] utilized a BW of 400 Hz which is used here as well. For any given filter bank, the BW is the same for all of the filters, which as we will see later, may not be ideal. A typical time-domain impulse response of the Gabor filter can be seen in the Figures 5.1 (b) and 5.2 (b). In fact both are the same filter with an (f_c) of 2,200 Hz and BW of 400 Hz.

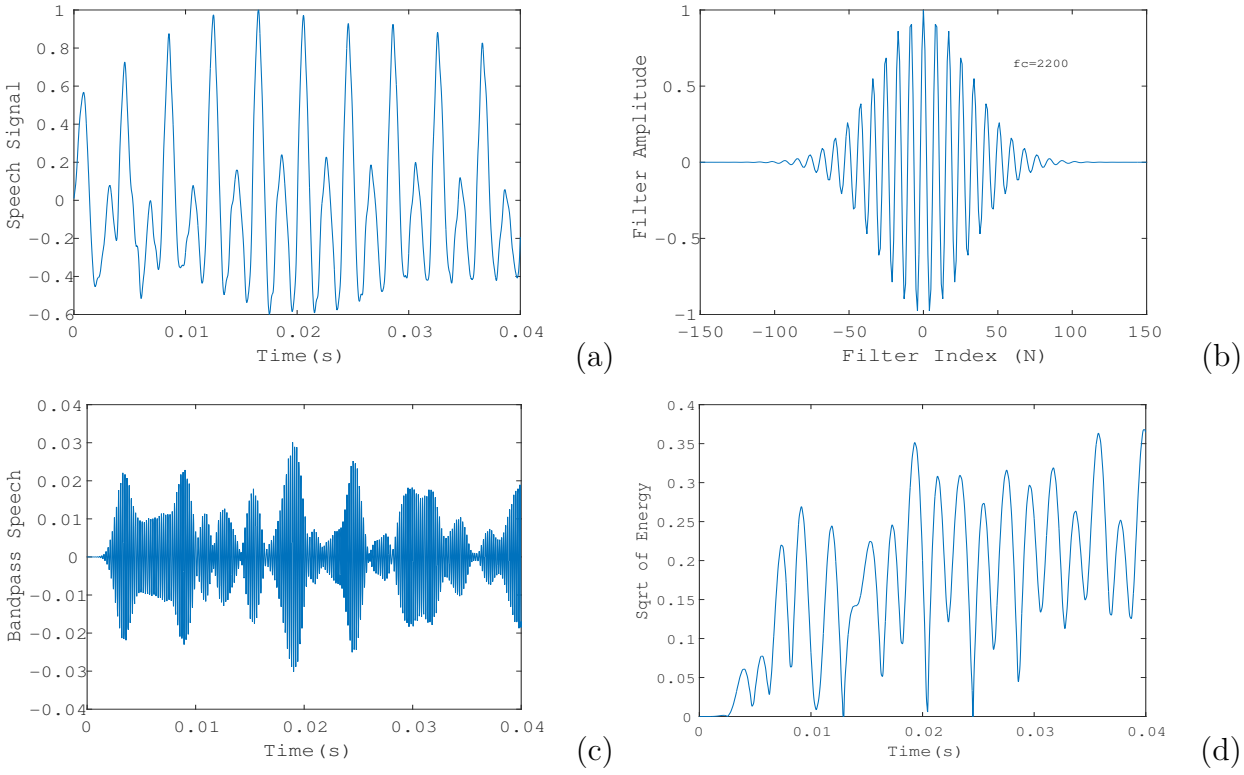


Figure 5.2: Processing steps of single-filterbank energy output for signal WOMENRS6. (a) time-truncated speech signal. (b) time-domain Gabor filter impulse response at bandpass fc of 2,200 Hz and BW of 400 Hz. (c) Gabor-filtered speech signal. (d) square root of Teager-Kaiser energy.

The equation for the continuous-time impulse response of the Gabor filter is:

$$h(t) = \exp(-\alpha^2 t^2)\cos(\omega_c t) \tag{5.1}$$

where t is in seconds, $\omega_c = 2\pi f_c$, and α controls BW through the relationship $BW_{rms} = \dfrac{\alpha}{\sqrt{2\pi}}$ as stated in [16]. The impulse response is then discretized by using the form:

$$h[n] = \exp(-b^2 n^2)\cos(\Omega_c n) \tag{5.2}$$

where

$n = T/t$ (T = sampling period in sec.), $b = \alpha T$, $\Omega_c = 2\pi f_c T$, and $-N \leq n \leq$

N, where N is chosen to truncate the Gaussian envelope essentially to zero; e.g., $\exp(-b^2N^2) \leq 10^{-5}$

Bandpass filtering is then performed by convolving $h[n]$ with the speech signal. It is important to strictly follow the constraints for N as the convolution operation will appear to time-shift and/or lengthen the tail of the speech signal if not properly adhered to. A check of the filter impulse response centered at 2,200 Hz, as well as all of the other center frequencies, shows that this constraint is always met. Filtered versions of the speech signals WOMENRS1 and WOMENRS6 can be seen in Figures 5.1 (c) and 5.2 (c). Note that the AM component of the speech signal is now more visible.

5.2 Intermediate processing steps

As previously stated, signal processing was performed on two signals from the ACP-CA database; WOMENRS1 (normal) and WOMENRS6' (nasal Level 6). The signals were recorded with 16-bit resolution and sampled at a rate of 44.1 kHz.

Figures 5.1 and 5.2 depict the intermediate processing results for the two signals at an f_c of 2,200 Hz. Since the original speech signal is swept from 200 Hz to 5,200 Hz in 20 Hz increments, there are 250 (number of Gabor filters) such sets of intermediate results that are generated. For the current set of plots, the filter BW is 400 Hz. A brief description of each subplot follows:

(a) Raw speech signal, time truncated. The original speech signals are up to 50 ms long but here are truncated to 20 - 40 ms since the speech pattern appears to repeat. The amplitudes have not been scaled or normalized.

(b) Impulse response of the Gabor filter at the band-pass center frequency of interest. For this set of data the center frequency (f_c) is 2,200 Hz.

(c) Band-pass filtered signal in (a) by filter in (b). Note that the modulated components of the signal, particularly the AM component, are more apparent.

(d) The square root of the Teager-Kaiser energy produced by the TKEO algorithm. Before taking the square root, the output from EO is amplitude shifted by the smallest amount necessary to eliminate negative values.

5.3 Energygram and interpretation of results

The 'enerygram' is produced by combining all 250 energy profiles into a color map that is plotted as a function of center frequency vs. time. The color scheme from highest to lowest energy is dark red/red/orange/yellow/green/light blue/blue/dark blue.

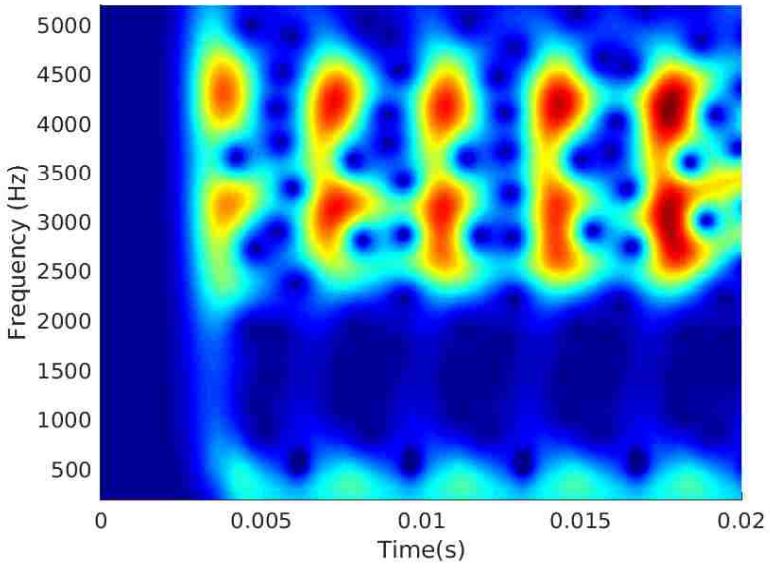
Enerygrams for the normal and nasal speech signals are shown in Figure 5.3. An examination of the normal signal (a) shows that the highest energies are in the region 2.5 kHz - 4.7 kHz with the two strongest energies at 3.0 kHz and 4.2 kHz. The energies at the lower frequencies are much weaker.

An examination of the nasal signal (b) shows that the highest energies are in the region 200 Hz - 700 Hz with the strongest energy at about 600 Hz. The energies at the higher frequencies are much lower or negligible.

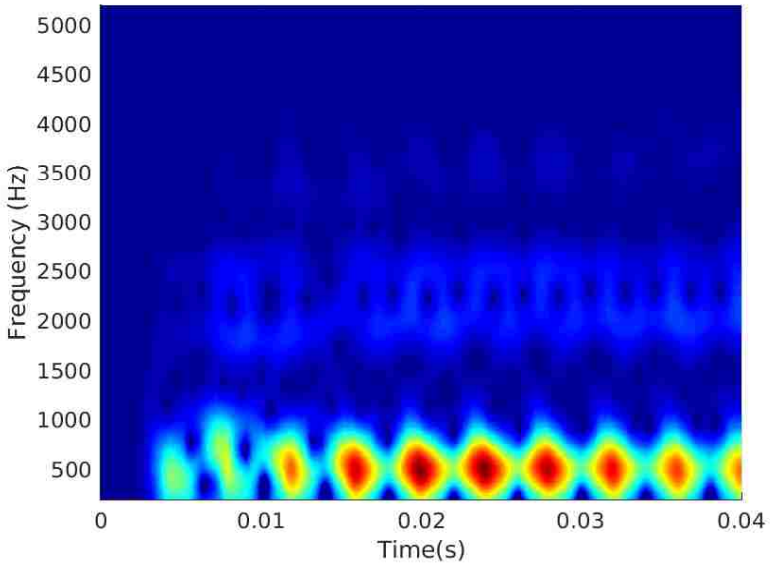
The differences in the energygrams are in general agreement with criteria set forth in [7] and in Chap 2.3, which states that for nasal speech, the spectral amplitude increases in the region between F1 and F2 and decreases in the F2 region. There are however some small discrepancies. For one, the spectral decrease for the nasal signal is actually closer to F3 than F2. Another is that there appears to be a F4 formant range that follows the same behaviour as the F3 formant. The reasons for these discrepancies may in part be due to the inability of a fixed- bandwidth Gabor

filter to successfully identify and separate the resonances of the original signal. This type of problem was pointed out in Chapter 1.

At this point, it appears that the energygram can be used to detect the *presence* of hypernasality but cannot determine the *level* of hypernasality. While detecting the presence of hypernasality is an important discovery, we still would like to be able to distinguish between the different levels. A promising method for determining the hypernasality level will be covered in the remaining two chapters.



(a)



(b)

Figure 5.3: (a) Energygram for normal voice signal WOMENRS1. (b) Energygram for nasal Level-6 nasal voice signal WOMENRS6. The nasal signal has higher spectral energy in the F1-F2 region and lower spectral energy in the F2 region. This is consistent with the criteria for hypernasality established in [5] and [7].

Chapter 6

Empirical Mode Decomposition

As can be seen in the previous chapters, all of the analysis methods for consistently detecting hypernasal speech have fallen short for various reasons. For the one-third octave band and LPC methods, the wideband nature of speech is the main culprit. For the more promising energy-based methods, the main culprit is the inability to spectrally isolate the resonant frequencies. If this were possible, not only would the energy-based methods show clearer results but there would be reason to revisit the LPC methods as was done in [27].

In this chapter, we investigate a relatively new method for waveform decomposition known as the EMD. Its success lies in the fact that it is an adaptive type of decomposition fully intended to be used for non-linear, non-stationary, and wide band signals.

6.1 The EMD Concept

Many types of signal decompositions (transformations), like for instance the Fourier transform, rely on a fixed set of orthogonal basis functions to perform the transforma-

tion. In the case of the Fourier transform, the basis functions are decided beforehand and remain fixed through the transform process without taking into account the nature of the signal being analyzed. Also, since the transform computes constant values of amplitude and frequency over a fixed time frame, it is only reliable for data that is of a stationary nature. In short, because it utilizes fixed basis functions, the Fourier transform is not an adaptive type of transform.

The wavelet transform, which is in many ways similar to the Fourier transform, is superior for transforming data that is non-stationary. This is because the basis functions can be altered before analysis to suit the nature of the data being analyzed. Keep in mind that once the basis functions are selected, they remain fixed for a given process. Unlike the Fourier transform however, the basis functions can be adjusted and the process can repeated until a suitable result is obtained. Thus the wavelet transform would appear to be a solution to our decomposition problem. The only problem here is that the choice of basis functions are a judgement call and can produce erratic results. In short, the wavelet transform is a pseudo-adaptive transform which requires a great amount of expertise to apply successfully.

For speech signals, the EMD is far superior to the Fourier or wavelet transforms, in that it is not based on a set of predetermined analytic basis functions. Instead, the basis functions for the algorithm are derived adaptively from the data signal itself. This makes it ideal for non-linear and non-stationary signals.

The EMD algorithm works by employing a sifting process that decomposes the input signal into constituent sub-signals called IMFs (Intrinsic Mode Functions). The IMFs are oscillatory in nature and indeed resemble AM/FM signals, which are consistent with our chosen speech model. By definition, the IMF must meet the following 2 criteria:

(1) the number of total extrema (min and max) and number of zero-crossings are the same or differ by 1, i.e., there is only 1 zero-crossing between 2 consecutive

extrema.

(2) at any point in time, the mean of the local maximum envelope and local minimum envelope is zero.

In meeting the above criteria, we are assured that the IMFs will be well-behaved, sinusoidal-like signals which are similar in nature to the resonances we seek. A summation of all of the IMFs will reproduce the original voice signal. This is analogous to summing the sinusoids from a Fourier transform to form the original signal before transformation.

For a typical voice signal, the EMD will produce from 7 to 12 true IMFs, and a residual IMF which is discarded as it represents the DC level of the signal.

6.2 The Basic EMD Algorithm

Before addressing the specifics of the EMD algorithm, we note that due to the adaptive nature of the EMD, there is no analytical or formal proof of the algorithm. It is a truly empirical algorithm (not derived or supported by mathematical theory) and the resulting IMFs are indeed empirical functions. However the algorithm has been tested on a large class of composed and real waveforms [22, 23, 24, 26, 27] with unambiguous results.

Figure 6.1 shows the beginning steps in the formation of an IMF. The signal for analysis is the damped sinusoid in solid blue. The positive and negative peak envelopes are estimated by fitting a cubic spline to the positive and negative peaks of the signal. The mean envelope (dashed gray line) is calculated by taking the average of the positive and negative peak envelopes. The mean envelope is subtracted from the signal and is tested by the two criteria stated in the previous section. If the criteria are met, then the first IMF has been calculated. If the criteria are not met,

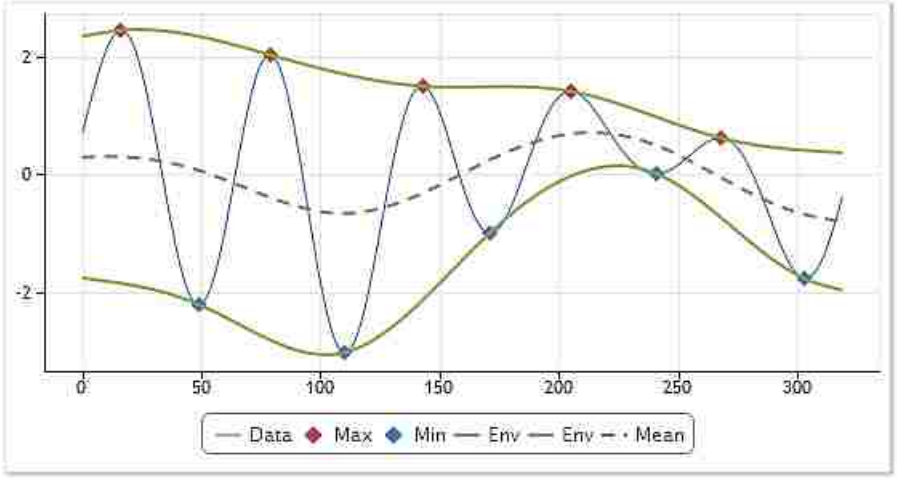


Figure 6.1: Basic EMD process for a damped sinusoid signal [31]. The positive and negative peak envelopes are estimated with cubic spline fit. The envelopes are then averaged and the mean envelope is subtracted from the original signal. The modified signal is tested by the IMF criteria.

the positive and negative envelopes are computed for the new waveform and the process repeats. This process is the *sifting* part of the algorithm. When the criteria is met for forming the first IMF, the IMF is subtracted from the initial signal and a new sifting process is started for the second IMF. This process is repeated until all of the IMFs and the residual have been produced.

A pseudocode representation for these steps similar to that in [24] is as follows:

$r_0(t) = X_0(t)$
%Sifting loop
for i=1:j
$h_0 = r_{k-1}$
extract local minima/maxima of $h_{j-1}(t)$
obtain envelopes $EMIN_{j-1}(t)$ and $EMAX_{j-1}(t)$
compute mean envelope $m_{j-1}(t) = [EMIN_{j-1}(t) + EMAX_{j-1}(t)]/2$
$h_j(t) = h_{j-1}(t) - m_{j-1}(t)$
if (Is IMF?)
YES. Extract kth IMF $d_k(t) = h_j(t)$
exit loop
else
NO. j=j+1
end

```
        end
r_k(t) = r_{k-1}(t) − d_k(t)
        if (at least 2 extrema?)
                jump to sifting loop
        else
                Exit loop. EMD complete
        end
end
```

Figure 6.2 shows the full EMD for an arbitrary waveform which consists of 4 IMFs and the residual. We note that the IMFs are oscillatory as expected, that they decrease in frequency with each new IMF, and that the residual does indeed track the DC level of the signal. In (a), the IMFs are plotted on the same vertical scale as the arbitrary waveform and in (b) they are auto-scaled so that the details can be seen. It is important to note that in (a), the IMFs span similar vertical ranges. This is a desirable feature as it indicates that there are no disparate modes in the original signal [22]. For speech signals, this will often not be the case and we will need a way to deal with it. This issue is dealt with in the next section.

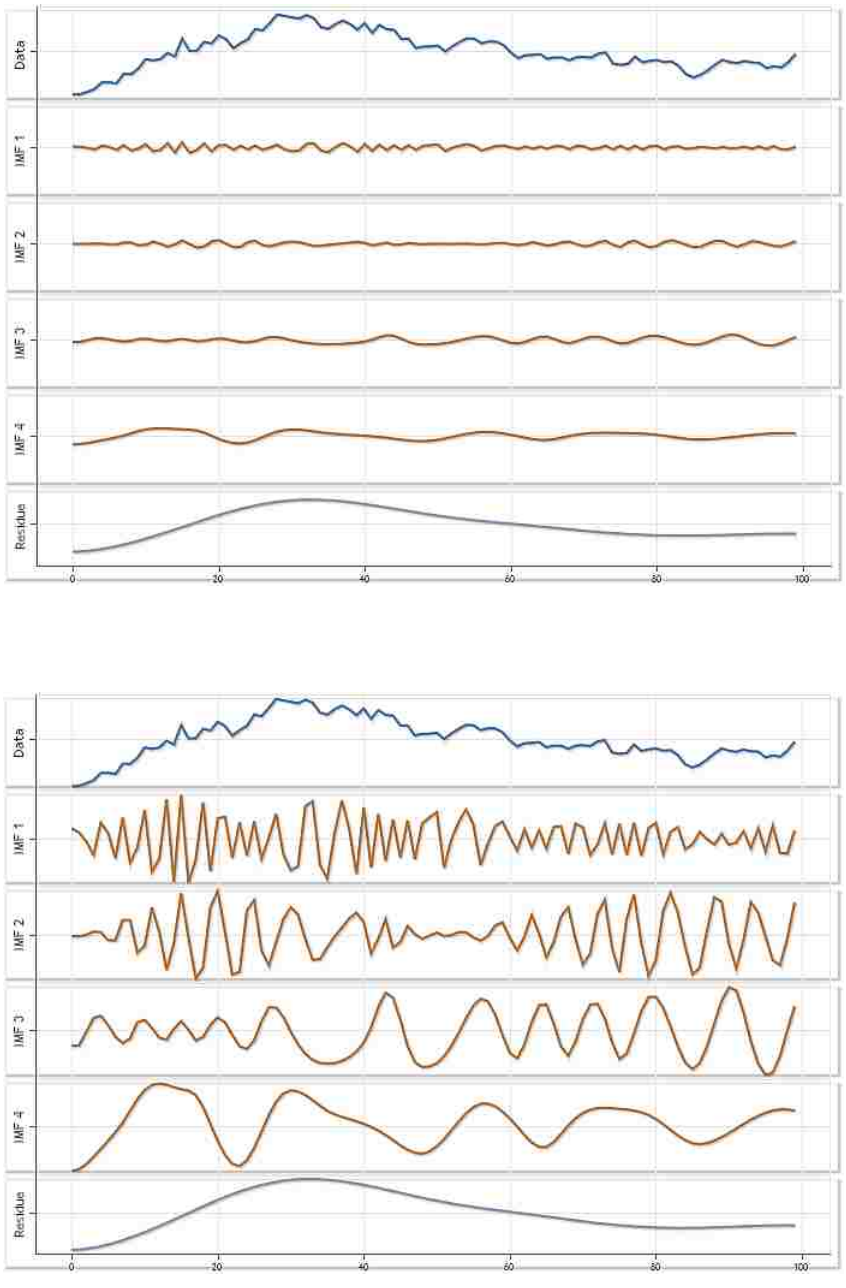


Figure 6.2: EMD of arbitrary data [31]. (a) Time plot of arbitrary data, 4 IMFs, and residual. (b) Same as (a) but with vertical auto-scaling of IMFs to see waveform details

6.3 Evolution of the basic EMD Algorithm

In its original form [23], the EMD is well-equipped to handle many classes of signals. As long as the signals are mainly oscillatory in nature, and do not have disparate modes, this version of EMD serves as a reliable mode separator [26]. However, for speech signals, especially the hypernasal ones, this will often not be the case as formants are often close to each other and as their bandwidths change, there can be spectral overlap. In these cases, modal behaviour that would be visible on one IMF may not be detected on a different IMF where the detail could be lost because because of the 'sensitivity' of the new IMF. Such a problem is known as mode mixing.

To deal with the mode-mixing problem, the basic EMD algorithm has been improved in several iterations. It has evolved from EMD, to EEMD, to Complementary EEMD (which was short lived), to CEEMDAN, and finally to the Improved CEEMDAN, which for this thesis, is assigned the name CEEMDAN-2014. The developmental details are well documented in [24]. In this section, only a brief overview of the evolutionary process is given. The main purpose here is to explain why the newest version, CEEMDAN-2014, was chosen for this work.

EMD - original algorithm developed by Huang in [30]

EEMD - Ensemble EMD - addresses the mode mixing problem by adding white Gaussian noise to copies of the original signal (ensemble), performing the decomposition, and then averaging the results. The downsides are: (a) the addition of residual noise to the recomposed signal and (b) the production of a different number of modes making final averaging difficult.

Complementary EEMD - attempts to deal with (a) above by using complementary (added and subtracted) pairs of noise but the completeness property cannot be verified and (b) is still present.

CEEMDAN - Complete EEMD w/ Adaptive Noise - solves the noise-in-

reconstruction problem and the different number of modes problem of the EEMD. Downsides are (c) some residual noise is present in some IMFs and (d) signal information appears in later IMFs, making earlier IMFs (the first 2 or 3) erroneous or spurious.

CEEMDAN-2014 - referred to as the 'Improved CEEMDAN' approach in [24]. Resolves issues (c) and (d) above but the number of the IMF's (modes) varies as in EEMD. Produces the least number of IMFs.

Chapter 7

EMD/Teager-Kaiser Energy Analysis of Hypernasal Voice Signals

We are now at the crux of the thesis where the strongest tools from the previous chapters are combined to define a strong marker for hypernasality. At this point, it should be apparent that the TKEO is the most suitable tool (ideal for non-linear, non-stationary and wideband data) for energy analysis of speech signals and that the EMD is the most suitable method (non-stationary, adaptive, oscillatory) for extracting resonances on which the energy operator can be applied. We shall now see how the hybrid EMD/TKEO approach works in practice.

In this chapter, the EMD is applied to normal and nasal voice signals from the ACP-CA database. A set of energy metrics (η_1 and η_2) is then derived from the IMFs via the TKEO that allows for, as we shall see, the delineation of the different hypernasal levels. Finally, the energy metrics plotted against the nasal levels stated in the database define a trend that links the energy to the hypernasality level .

7.1 Voice Signals from the Cleft-Palate Database

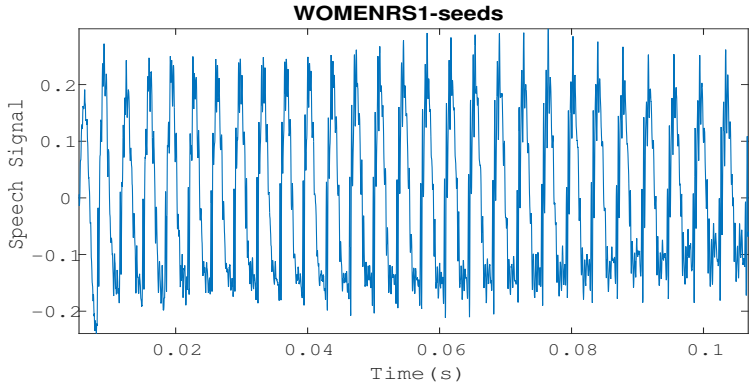
The voice samples used for analysis were taken from the *American Cleft Palate - Craniofacial Assoc.*, or ACP-CA database [28]. This database is the most appropriate one found to-date as the subjects are cleft palate patients and exhibit various levels of hypernasality [29]. The database consists of 6 to 8 voice samples each for men, women, and children. The nasal levels of the subjects were determined on basis of perceptual evaluation by clinician(s), and range from 1 to 8, with 1 indicating normal speech and 8 indicating extreme hypernasality. Only the vowel /i/ was chosen for analysis. At this time, the nasalance scores for the voice samples are not available.

Two sets of voice samples (from the SHS lab and UNM-DSP lab) were recorded but were not analyzed. The reason for this was because these voice samples were those of normal speakers *feigning* hypernasality. Our presumption is that the mechanisms that produce true hypernasal speech *could* be fundamentally different from those for feigned hypernasal speech. Until this presumption can be investigated (see Chap. 8.2), conclusions will only be drawn from the ACP-CA data set.

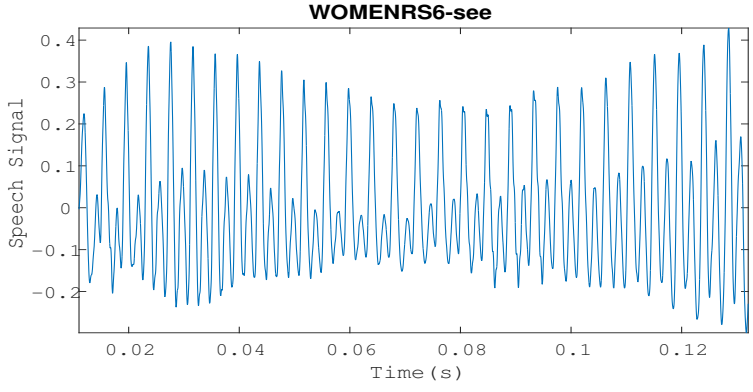
7.2 Energy Metrics and Pseudo-Classification

Three voice samples each for women, men, and children (9 total) were analyzed. These were:

(1) Women: Levels 1, 3, 6
(2) Men: Levels 1, 4, 6
(3) Child: Levels 1, 5, 6



(a)



(b)

Figure 7.1: Voice samples of vowel /i/ from ACP-CA database [29]. (a) Normal voice signal 'WOMENRS1'. Vowel is extracted from utterance "seeds". (b) Nasal voice signal 'WOMENRS6. Vowel is extracted from the utterance "see".

Voice samples where the vowel /i/ was clearly distinguishable were chosen. Care was taken to avoid words where the vowel was proceeded or followed by a nasal consonant such as /n/ or /m/. This was so that additional nasal components would not be introduced into the voice samples.

Figure 7.1 shows speech samples of the vowel /i/ for the signals WOMENRS1 and WOMENRS6, which as the names indicate are nasal Level-1 (normal) and hypernasal Level-6. The full length of the vowel (up to 30 pitch periods) was chosen for analysis.

Figure 7.2 shows the 'relevant' IMFs that were produced for each of the signals from Figure 7.1. Here, a 'relevant' IMF is one that contributes significant energy ($\geq 1\%$) to the total energy spectrum (Energy computations are done later). As can be seen, the IMFs look as they should: they are oscillatory and generally look like AM/FM signals. Note that the frequency content tends to decrease with higher numbered IMFs.

Spectrograms for each IMF are shown in Figures 7.3 and 7.4. The spectrograms verify that the lower-numbered IMFs retain more of the original signal's high-frequency content while the higher-numbered IMFs retain more of the low-frequency content. Later, the spectrograms will be used with the energy metrics to establish classification criteria for the different hypernasality levels.

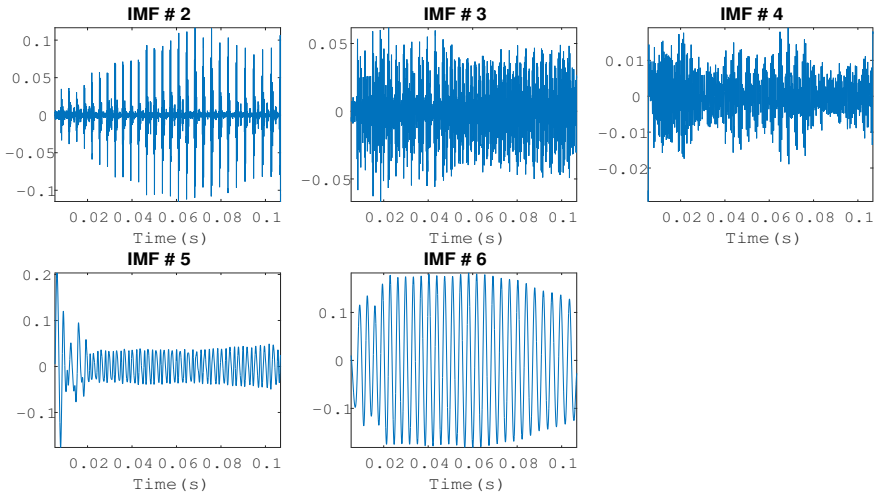
The energies for each IMF were then computed with the TKEO. (IMF energy plots are shown in Appendix A). Energy metrics were then derived for the higher and lower frequency bands. For this work, the higher frequency bands are those above 1 kHz ($f_s = 44.1$ kHz) and the lower frequency bands are those below 1 kHz ($f_s = 44.1$ kHz).

The energy metric for the upper frequency bands is given by:

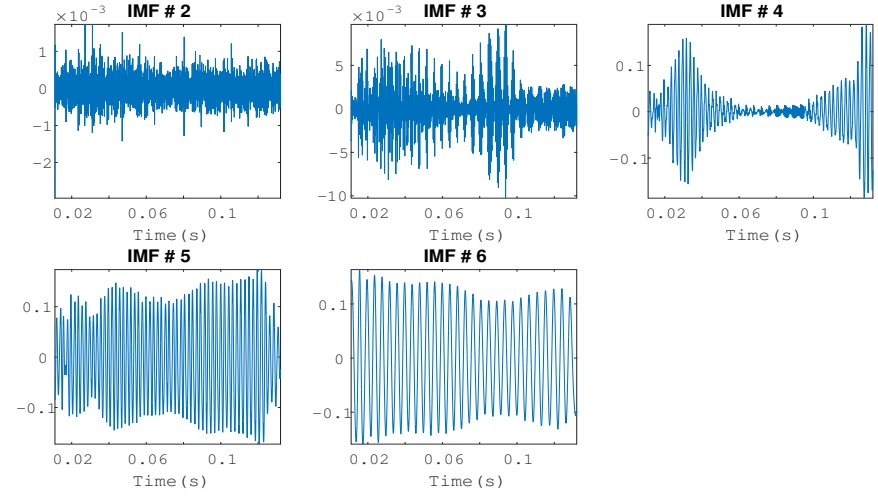
$$\eta_1[k] = \frac{\sum_{i=1}^{k} \tilde{\psi}\left(m_i[n]\right)}{\sum_{i=1}^{n} \tilde{\psi}\left(m_i[n]\right)} \tag{7.1}$$

The energy metric for the lower frequency bands is given by:

$$\eta_2[k] = \frac{\sum_{i=k}^{n} \tilde{\psi}\left(m_i[n]\right)}{\sum_{i=1}^{n} \tilde{\psi}\left(m_i[n]\right)} \tag{7.2}$$

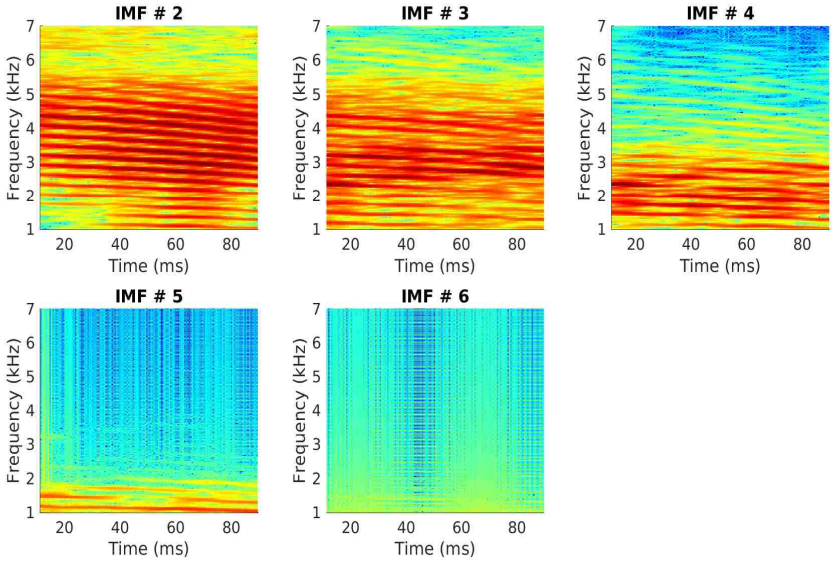


(a)

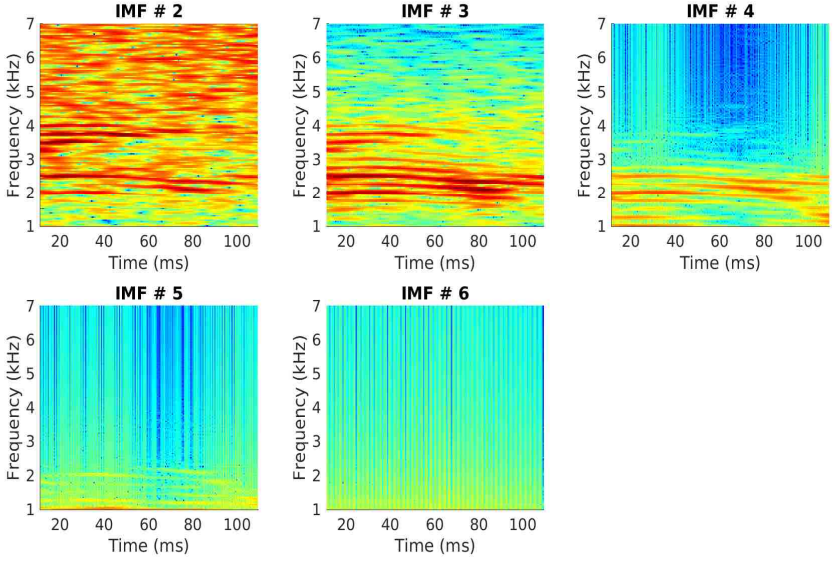


(b)

Figure 7.2: Selected IMFs of voice samples from Figure 7.1. (a) Normal voice signal of utterance "seeds" from speaker 'WOMENRS1'. (b) Nasal voice signal of utterance "see" from speaker 'WOMENRS6'. IMFs were produced by the CEEMDAN-2014 algorithm. The lower-numbered IMFs retain more of the original signal's high-frequency content while the higher-numbered IMFs retain more of the low-frequency content. Only 5 (displayed) of the 12 produced IMFs for each signal contribute significantly to the total signal energy. EMD decomposition into IMFs permits a clearer analysis of local signal oscillations by breaking down the voice signal into modular amplitude and frequency components.

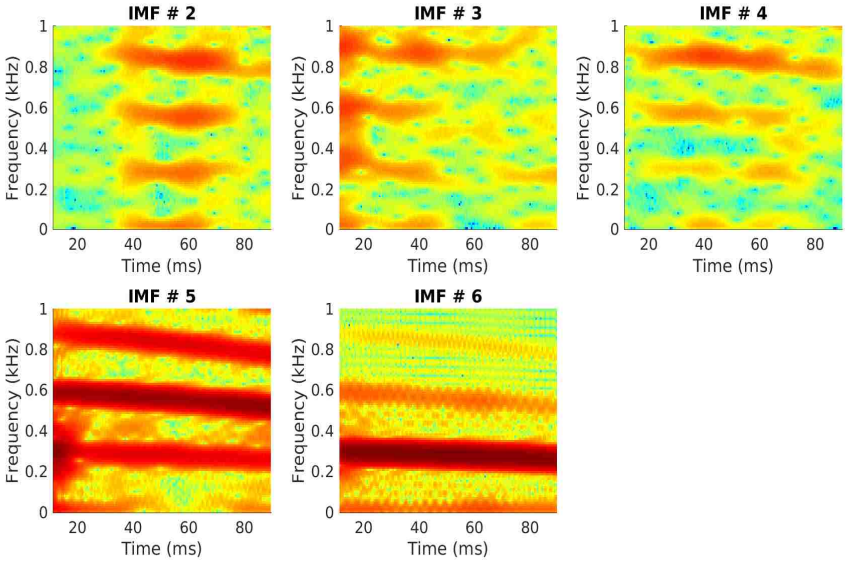


(a)

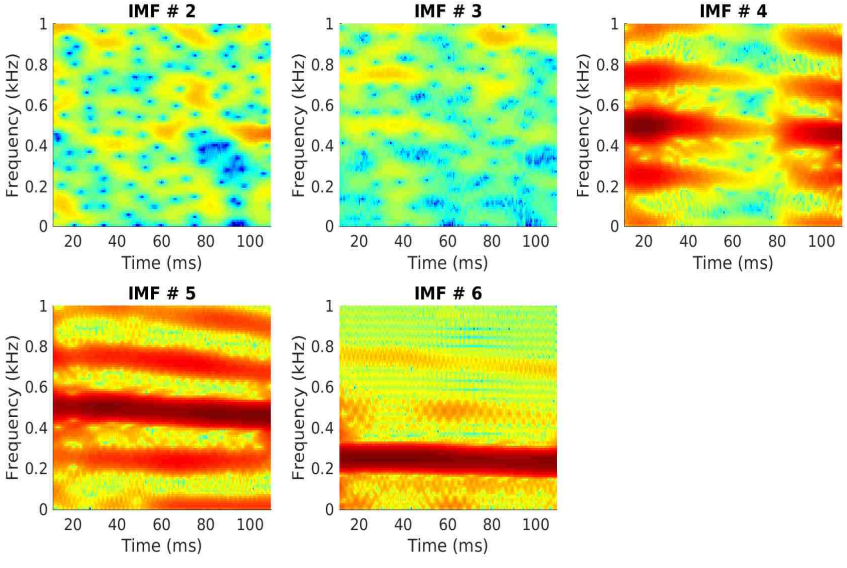


(b)

Figure 7.3: Spectrograms of IMFs from Figure 7.2. (a) 1-7 kHz range for normal voice signal. (b) 1-7 kHz range for nasal voice signal. Comparing IMFs #3 and #4 from (a) and (b), a much weaker high-frequency content is observed above 3 kHz for the nasal signal, indicating the formation of an anti-resonance or a range of anti-resonances. This is the first marker for hypernasality.



(a)



(b)

Figure 7.4: Spectrograms of IMFs from Figure 7.2. (a) 0-1 kHz range for normal voice signal. (b) 0-1 kHz range for nasal voice signal. Comparing IMF #4 from (a) and (b), a much stronger low-frequency content is observed in the range 200 Hz - 1 kHz for the nasal signal indicating the formation of a resonance or range of resonances. This is the second marker for hypernasality.

For equations 7.1 and 7.2, η denotes the set of energy metrics, $m_i[n]$ is the i_{th} IMF in a set of IMFs for a particular signal, $\tilde{\psi}$ is the medium energy after applying the TKEO, k is a counter limit whose value is determined by the classification criteria (1 kHz), and n is the total number of IMFs for a particular signal.

In short, the energy metric η is a percentage measure of the energy that is contained in a partial sum of IMFs. For equation 7.1, the summation in the numerator begins at the lowest-numbered IMFs, which contain the high-frequency content, and sums forward. For equation 7.2, the summation in the numerator begins at the highest-numbered IMFs, which contain the lowest-frequency content, and sums backward. For both equations, the partial energy sums are computed as ratios to the median total energy, thus producing an energy percentage.

Energy metrics (η_1 and η_2 values) for the three women's voices are shown in Table 7.1. For completeness, the η values are calculated out to $k = n$. This will result in all η_1 values ending at 1 and all η_2 values beginning at 1.

The bolded entries of Table 7.1 represent the η values that meet the 1 kHz classification criteria and are used to judge the level of hypernasality. The frequency content for each IMF can be roughly judged from the spectrograms in Figures 7.3 and 7.4, but they were more precisely determined using the FFT.

In words, Table 7.1 (and the following tables) can be interpreted as follows: "As nasality increases from Level-1, to Level-4, to Level 6, the energies in the upper frequency bands (η_1) correspondingly decrease from 76%, to 21%, to 9.7%. For the same nasality levels, the energies in the lower frequency bands (η_2) correspondingly increase from 25%, to 71%, to 99%."

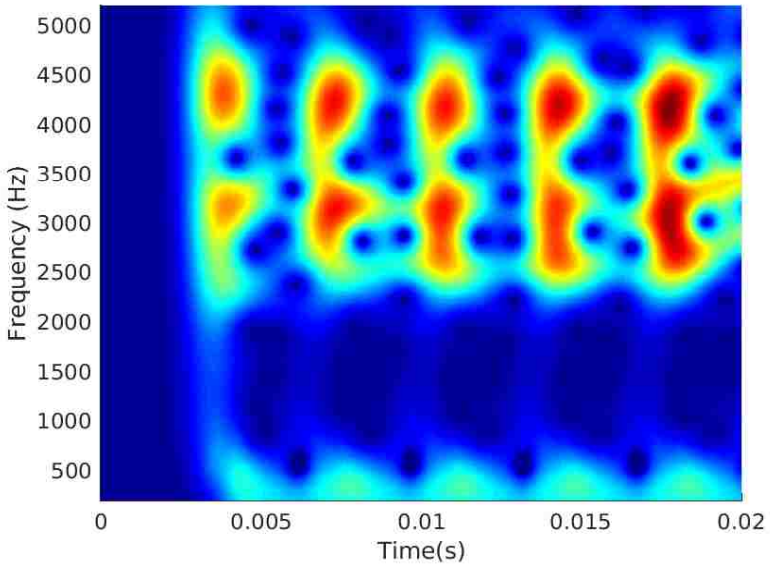
At this point, the energygrams from Figure 5.3 are redisplayed in Figure 7.5 so that we can compare them to the WOMEN energy metrics just obtained. On comparing (a) to (b), we note that the decrease of energy intensities at high frequencies of 3.0 kHz and 4.3 kHz are consistent with the decreasing η_1 values from Table 7.1.

IMF	η_1:normal	η_2:normal	η_1:nasal3	η_2:nasal3	η_1:nasal6	η_2:nasal6
1	0.0069	1.0000	0.0027	1.0000	0.0010	1.0000
2	0.0695	0.9865	0.0256	0.9933	0.0017	0.9983
3	0.7294	0.8683	**0.2069**	0.9731	0.0072	0.9979
4	**0.7590**	0.2728	0.2069	**0.7123**	**0.0972**	**0.9889**
5	0.8088	**0.2449**	0.8085	0.6752	0.8302	0.8261
6	1.0000	0.2032	1.0000	0.1467	0.9999	0.1615
7	1.0000	0.0000	1.0000	0.0000	1.0000	0.0001
8	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000
9	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000
10	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000

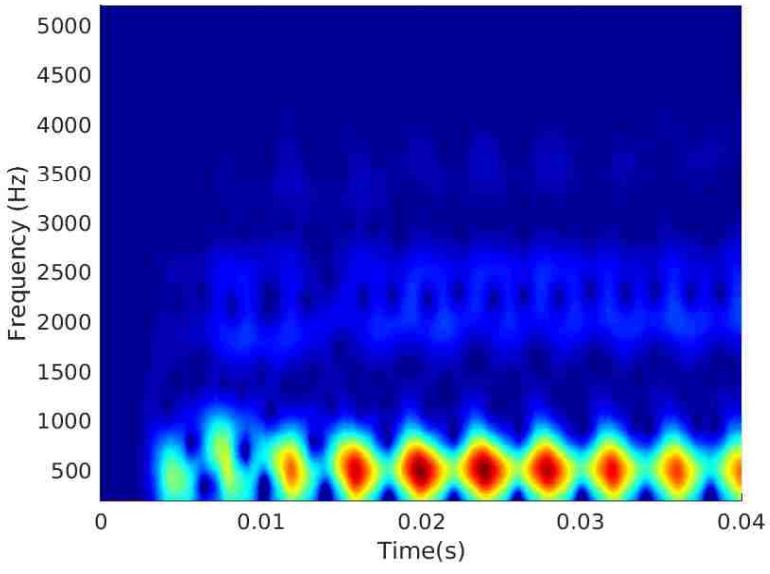
Table 7.1: Energy metrics for normal voice signal 'WOMENRS1', Level-3 nasal voice signal 'WOMENRS3', and Level-6 nasal voice signal 'WOMENRS6' from [29]. The bolded η_1s represent the maximum η_1 values that meet the classification criteria of containing frequency content >1 kHz. The bolded η_2s represent the maximum η_2 values that meet the classification criteria of containing frequency content <1kHz. For this and other normal/nasal comparisons from the ACP-CA database, the nasal η_1s are typically lower than the normal η_1 while the nasal $\eta_2 s$ are typically higher than the normal η_2. The consistency of this dual comparison indicates the presence of a strong marker for hypernasality.

We can also see that the increase of energy intensities at a low frequency of 600 Hz is consistent with the increasing η_2 values also from Table 7.1. As was noted in Sec 5.5.3, the energygram is a marvellous tool for detecting nasality, but not necessarily useful for quantifying nasality. It is reintroduced here merely to demonstrate that the energy metrics are a step in the right direction towards delineating between different hypernasality levels.

Energy metrics were similarly derived for men and children. These results are shown in Tables 7.2 and 7.3. The dual trends of decreasing η_1 and increasing η_2 as a function of increased hypernasality are observed here as well.



(a)



(b)

Figure 7.5: (a) Energygram for the normal voice signal WOMENRS1. (b) Energygram for the nasal Level-6 voice signal WOMENRS6. The decrease of energy intensities at high frequencies of 3.0 kHz and 4.3 kHz are consistent with the decreasing η_1 values from Table 7.1. The increase of energy intensities at a low frequency of 600 Hz is consistent with the increasing η_2 values also from Table 7.1.

IMF	η_1:normal	η_2:normal	η_1:nasal4	η_2:nasal4	η_1:nasal7	η_2:nasal7
1	0.0510	1.0000	0.0412	1.0000	0.0149	1.0000
2	0.2714	0.9206	0.0970	0.9274	0.0266	0.9769
3	0.7494	0.5935	0.1487	0.8683	0.0478	0.9661
4	**0.8343**	0.2220	**0.2470**	0.7488	**0.0577**	**0.9393**
5	0.9774	**0.1332**	0.8573	**0.5679**	0.0703	0.9131
6	0.9945	0.0237	0.9818	0.1456	0.6286	0.8835
7	1.0000	0.0051	0.9996	0.0150	0.9995	0.2233
8	1.0000	0.0000	1.0000	0.0001	1.0000	0.0015
9	1.0000	0.0000	1.0000	0.0000	1.0000	0.0001
10	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000
11	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000

Table 7.2: Energy metrics for normal voice signal 'MENRS1', Level-4 nasal voice signal 'MENRS4", and Level-7 nasal voice signal 'MENRS7' from [29]

IMF	η_1:normal	η_2:normal	η_1:nasal5	η_2:nasal5	η_1:nasal6	η_2:nasal6
1	0.0062	1.0000	0.0003	1.0000	0.1030	1.0000
2	0.7378	0.9883	0.0028	0.9995	0.3114	0.7636
3	0.9386	0.2476	0.0109	0.9925	0.4445	0.4334
4	**0.9489**	**0.0585**	0.0226	0.9771	**0.5304**	0.3766
5	0.9824	0.0514	**0.6603**	0.9517	0.6185	**0.3214**
6	1.0000	0.0114	0.9998	**0.2900**	0.9370	0.2796
7	1.0000	0.0000	0.9999	0.0002	0.9993	0.0342
8	1.0000	0.0000	1.0000	0.0001	1.0000	0.0007
9	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000
10	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000
11	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000

Table 7.3: Energy metrics for normal voice signal 'CHILDRS1', Level-5 nasal voice signal 'CHILDRS5', and Level-6 nasal voice signal 'CHILDRS6' from [29]

Figure 7.6 shows the η values for all 9 speech signals plotted as a function of nasal level. Of note here is the monotonically decreasing nature of η_1 with increasing hypernasality and the monotonically increasing nature of η_2 with increasing hypernasality. This trend is present amongst all of the chosen subjects and appears to be consistent across gender and age.

In attempting to curve-fit the different η sets, linear regressive power, exponential, and 2nd-order polynomial fits were tested. Intuitively, one would think that the energy level should level-off (approach zero slope) as the nasal levels approach a maximum. Therefore the best curvatures were judged to be convex-like for the η_1 fits and concave-like for the η_2 fits.

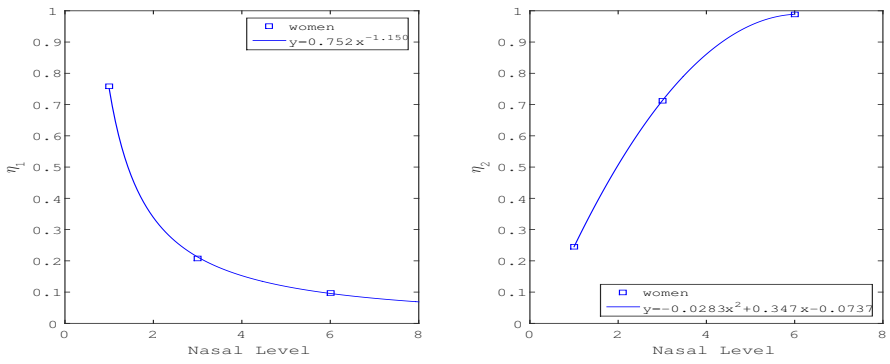
For the η_2 sets, the power and exponential fits did not work well and even exhibited slightly convex behaviour. Therefore all of the η_2 sets were fit to a 2nd-order polynomial which looks quite appropriate. However, it should be recalled that 3 data points form an overdetermined system for a 2nd-order polynomial and the fit will pass through all of the data points.

For the η_1 sets, the power fit works very well for women while an exponential fit looks better for the men. The best fit for the children is a 2nd-order polynomial, in spite of the wrong curvature.

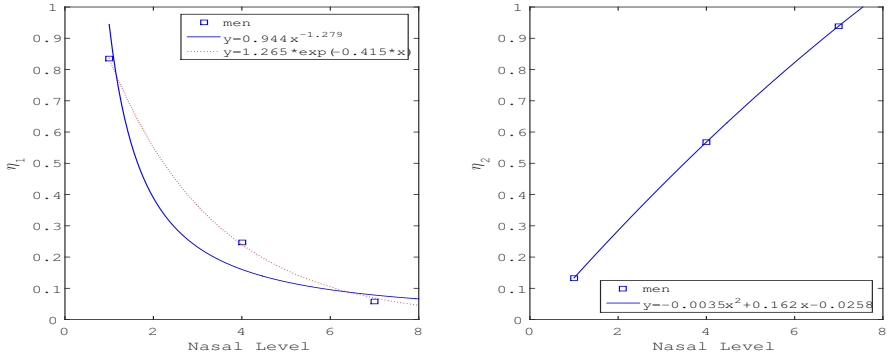
The slight ambiguities in the η_1 curves, especially for the children, expose a potential problem with the EMD/TKEO approach. Because the CEEMDAN-2014 algorithm minimizes the number of IMFs that are produced, there is not a high amount of resolution in the energy metrics. For instance, if we examine Table 7.3 and compare IMF #4 to IMF #5 for nasal Level 5 of η_1, we note that the energy metric jumps from 2% to 66%.

Earlier versions of the EMD algorithm, such as CEEMDAN that can produce as many as 17-20 IMFs for this type of data, would not fare much better as many of the IMFs would be throwaways [24].

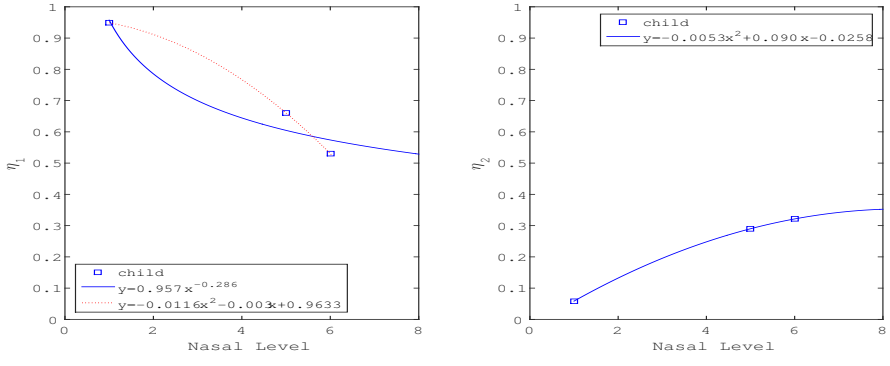
If we can accept that 1 of the 9 energy metrics is an outlier or falls within a reasonable statistical bound, the following hypothesis can be made:



(a)



(b)



(c)

Figure 7.6: η_1 and η_2 values vs. Nasal Level for voices from [29]: (a) Women. (b) Men. (c) Children. The nasality of subjects are determined on the basis of perceptual evaluation by a clinician and range from 1 to 8, with 1 indicating normal speech and 8 indicating extreme hypernasality. Of note here, is the monotonically decreasing nature of η_1 with increasing hypernasality and the monotonically increasing nature of η_2 with increasing hypernasality. The trend is consistent across both age and gender.

For a given classification criterion, the η_1 and η_2 energy metrics are able to delineate between different levels of hypernasal speech. For the vowel /i/, the energy metrics for the higher formants (F1-F2) decrease monotonically with increased hypernasality while the metrics for the lower formants (F1) increase monotonically with increased hypernasality.

We note that the best curve fits for η_1 are either power or exponential and the best curve fit for η_2 is a 2nd-order polynomial. We also note that the different curve fits might indicate that the formation of resonances and anti-resonances are different in nature. This is consistent with the fact that a resonance corresponds to constructive interference and an anti-resonance corresponds to destructive interference.

Chapter 8

Conclusions and Future Work

8.1 Conclusions

In this thesis, a novel hybrid EMD/TKEO approach was developed to address the inadequacies of existing approaches for hypernasal speech detection. The traditional methods of formant analysis were shown to have limitations in terms of their suitability for analyzing nasal and hypernasal speech signals. Furthermore, these methods were not able to discern between different levels of hypernasality. The hybrid EMD/TKEO approach, however, was able to produce a clear delineation between hypernasality levels when applied specifically to voice samples from the ACP-CA database.

At this time a few cautionary remarks must be made about the voice samples from the ACP-CA database; in listening to all of the voice samples, there appear to be some patients that have speech problems in addition to the hypernasality issues. For instance, the voice sample CHILDRS6 (used for this study), has an irregular voicing source, possibly indicating that the vocal folds are not vibrating synchronously. The voice sample MENRS7 (used for this study) has a drop of acoustic energy towards

the end of the vowel /i/.

There are also some voice samples that are compromised by the quality of the sound recording. For instance, the sample WOMENRS8 (not used for this study) has a high level of background noise, more than likely caused by the low level of 'sound' exiting through the mouth for such a high level of hypernasality. In the voice sample MENRS8 (not used for this study), there appears to be a typewriter in the background towards the end of the recording.

Such issues with this particular database will limit the scope of this study as we must assume that some of the energy metrics *could* be determined by factors outside of hypernasality. However, since the aforementioned issues are found to be prevalent in subjects with high hypernasality, the EMD/TKEO results are still valid, but perhaps only up to certain nasal levels, depending on subject group. Based on the irregularities found in the database, the energy metrics for Men and Children should perhaps be re-evaluated with voice samples taken at Nasal Levels 6 and 5 respectively. The energy metrics for Women are not in question as none of the noted irregular features were present in the chosen speech samples for this group. This may in fact be the main reason why the η_1 vs. Nasal Level plot for Women has the best curve fit.

8.2 Future Work

While the EMD/TKEO approach to hypernasal speech analysis looks promising, the results for this work were obtained from a fairly small set (9) of data samples. Future work would entail testing the method on larger and different types of speech databases. Future tasks would be to:

(1) Apply the EMD/TKEO to the remaining samples of ACP-CA database to see how well the results compare to the energy metrics established in Chapter 7.2.

The same type of analysis should also be done for the vowels /a/, /e/ and /o/ and compared with the results in [16], [20], and [27].

(2) Develop more stringent classification criteria for determining the energy metrics that determine the hypernasality levels. While a single threshold level of 1 kHz proved sufficient for this thesis, multiple threshold levels based on the traditional F1, F2, and F3 formant levels may prove optimal.

(3) Cross-validate the hypernasality results by using the EMD/TKEO to 'generate' a database of synthetic vowels with different levels of hypernasality. The levels of synthetic hypernasality would then be compared with the perceptual evaluation by a clinician.

(4) Compare different causes of nasality. This would entail analyzing the data from the SHS and UNM-DSP labs to determine if the mechanisms for true and feigned hypernasal speech are similar or different from each other.

(4) Apply the EMD/TKEO to voice samples recorded in the SHS lab where nasalance scores were collected with the voice data. Generate η vs. Nasalance Score plots and compare with the η vs. Nasal Level obtained in Fig. 7.6 to see how the trends compare. The η vs. Nasalance Score data should also be compared with the one-third octave band criteria established in [5].

Appendices

Appendix A

EMD Decompositions from ACP-CA Database

A.1 IMFs of analyzed signals

This Appendix shows plots IMFs, spectrograms, and energies that are not in the main body of the thesis.

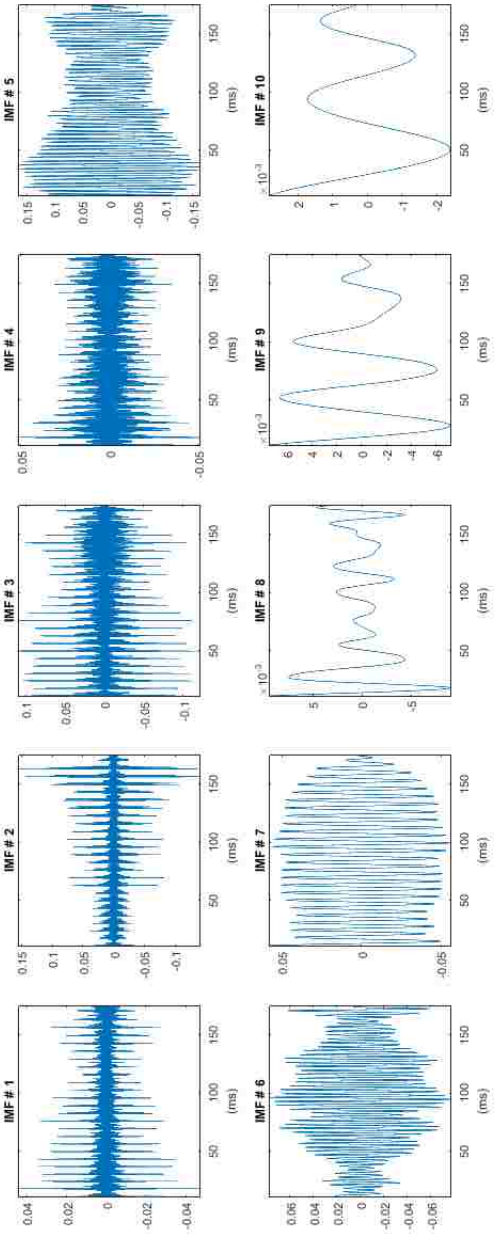


Figure A.1: IMFs for signal MENRS1-feed

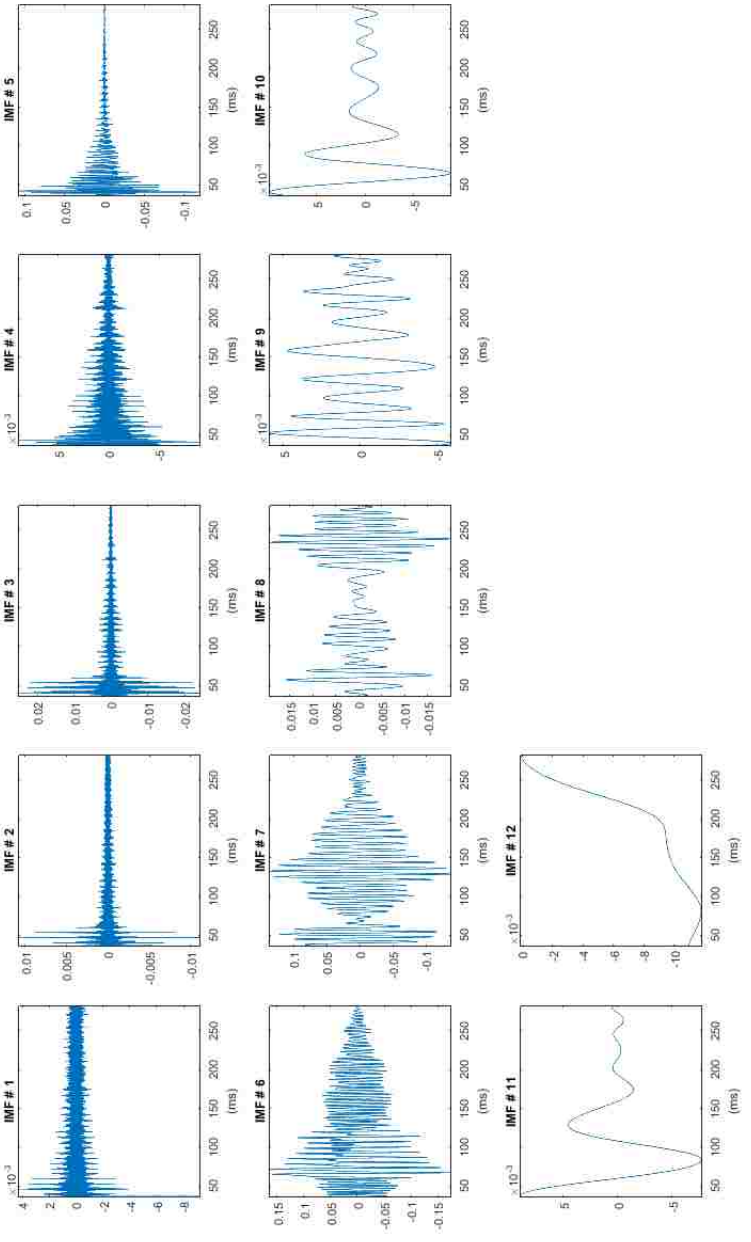


Figure A.2: IMFs for signal MENRS7-feet

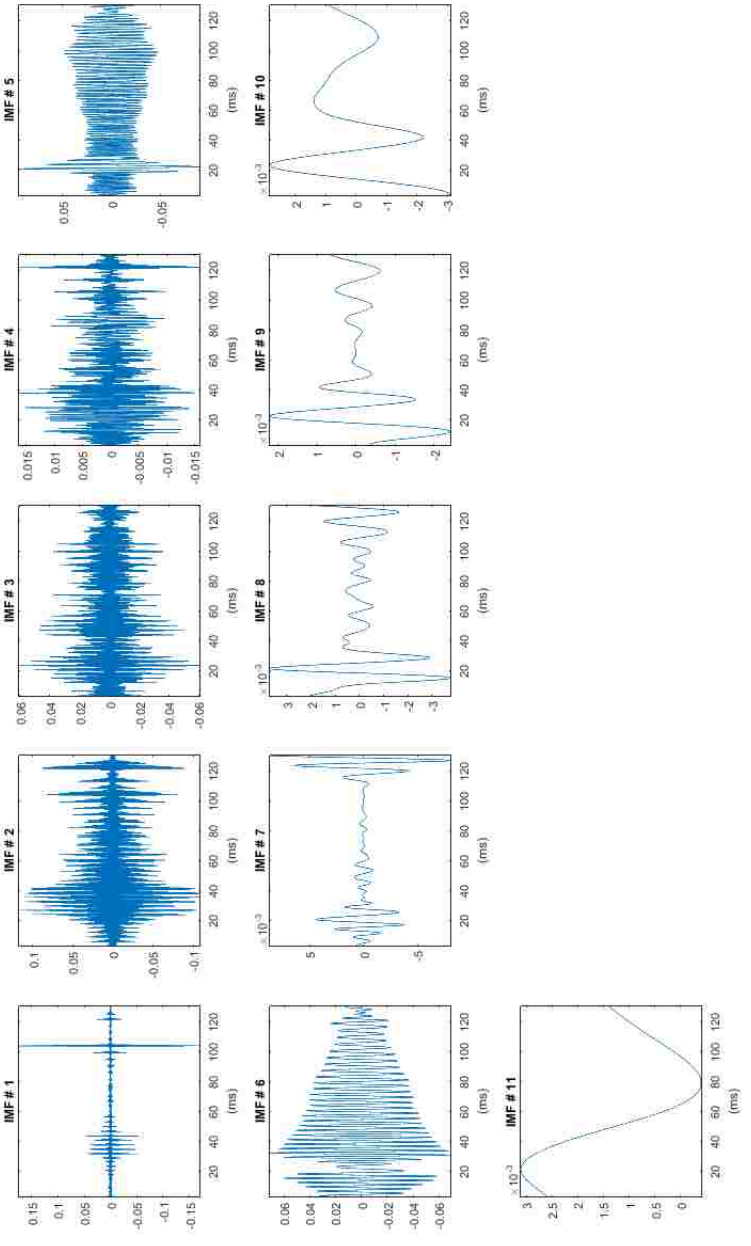


Figure A.3: IMFs for signal CHILDRS1-feet

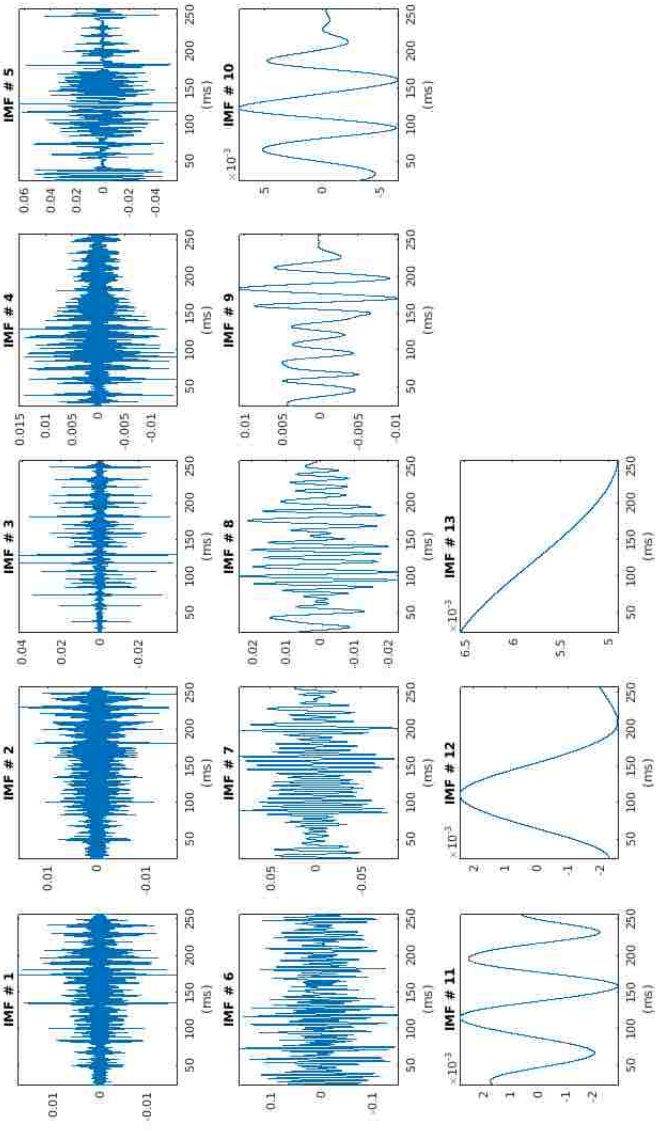


Figure A.4: IMFs for signal CHILDRS6-street

A.2 Spectrograms of IMFs for analyzed signals

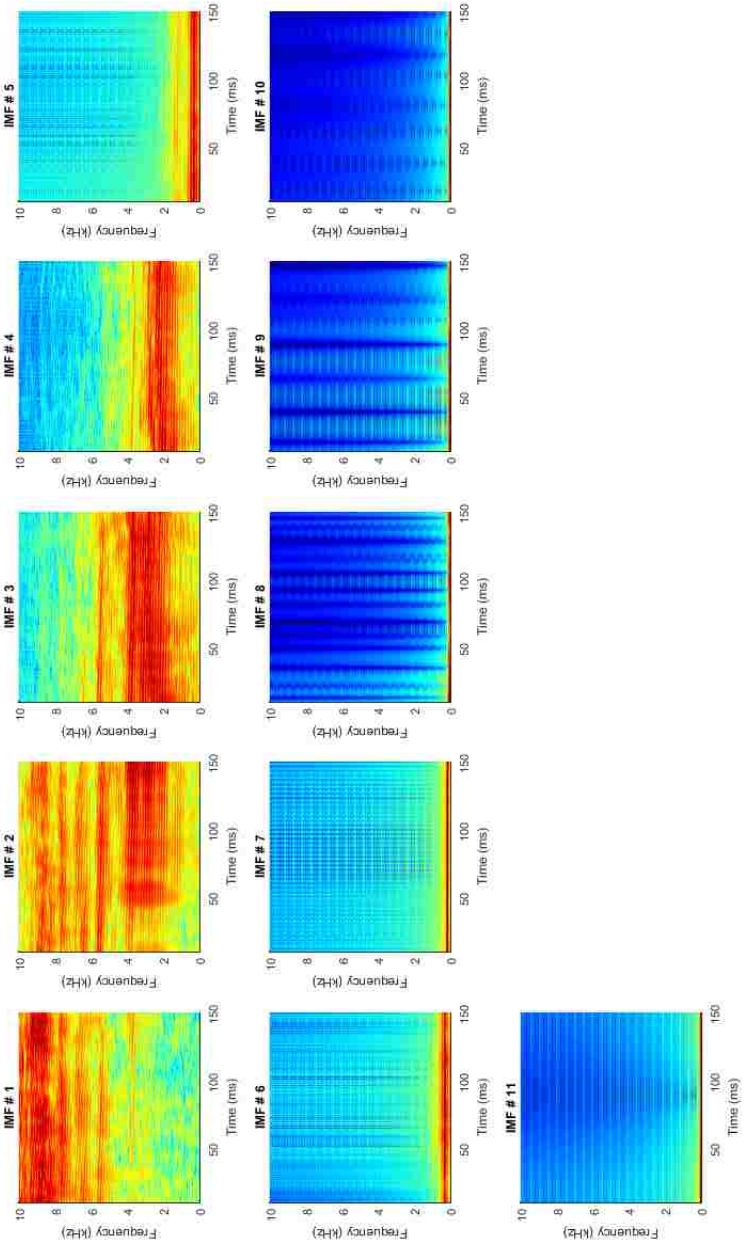


Figure A.5: Spectrograms of IMFs for signal MENRS1-feed. Range: 0 to 10 kHz

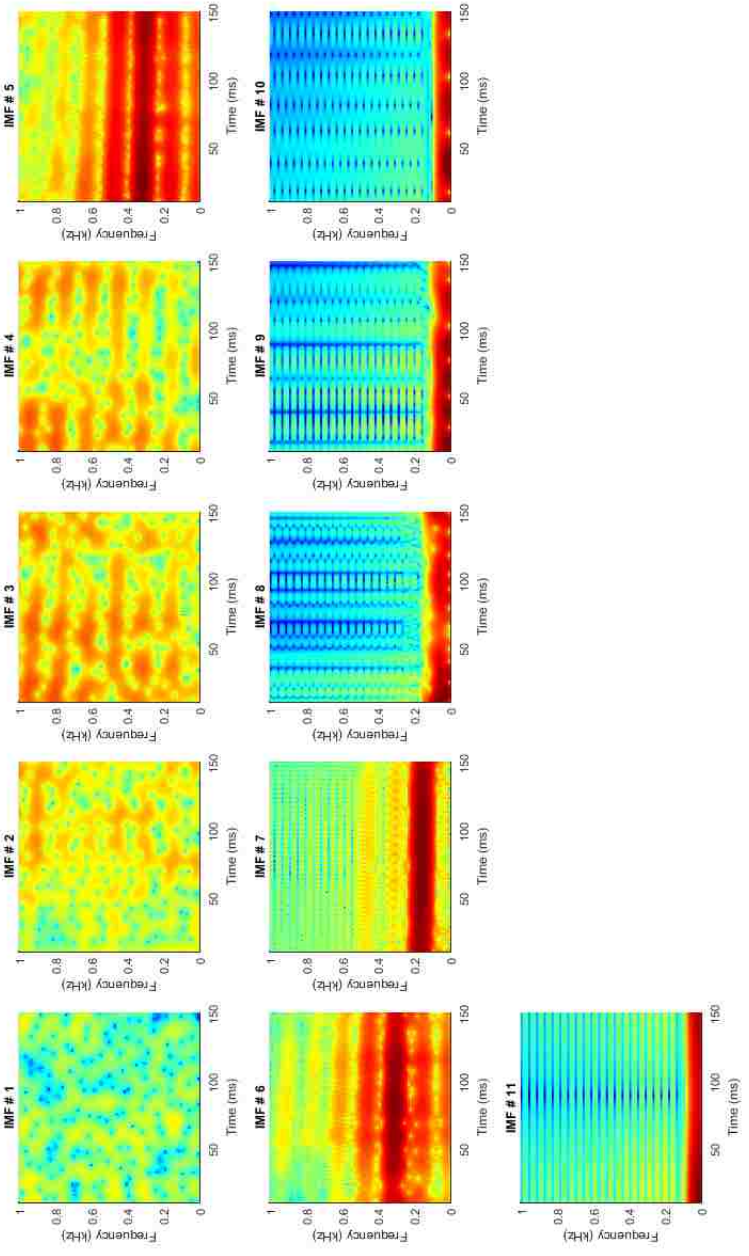


Figure A.6: Spectrograms of IMFs for signal MENRS1-see. Range: 0 to 1 kHz

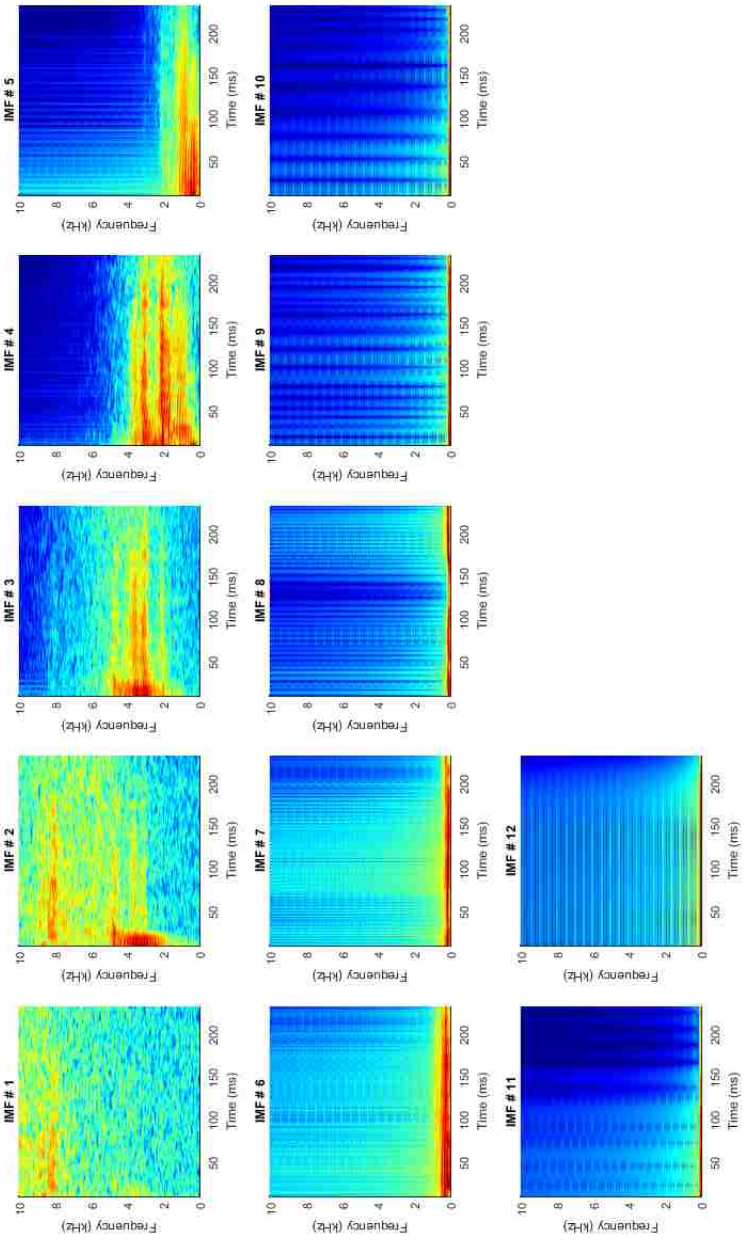


Figure A.7: Spectrograms of IMFs for signal MENRS7-feet. Range: 0 to 10 kHz

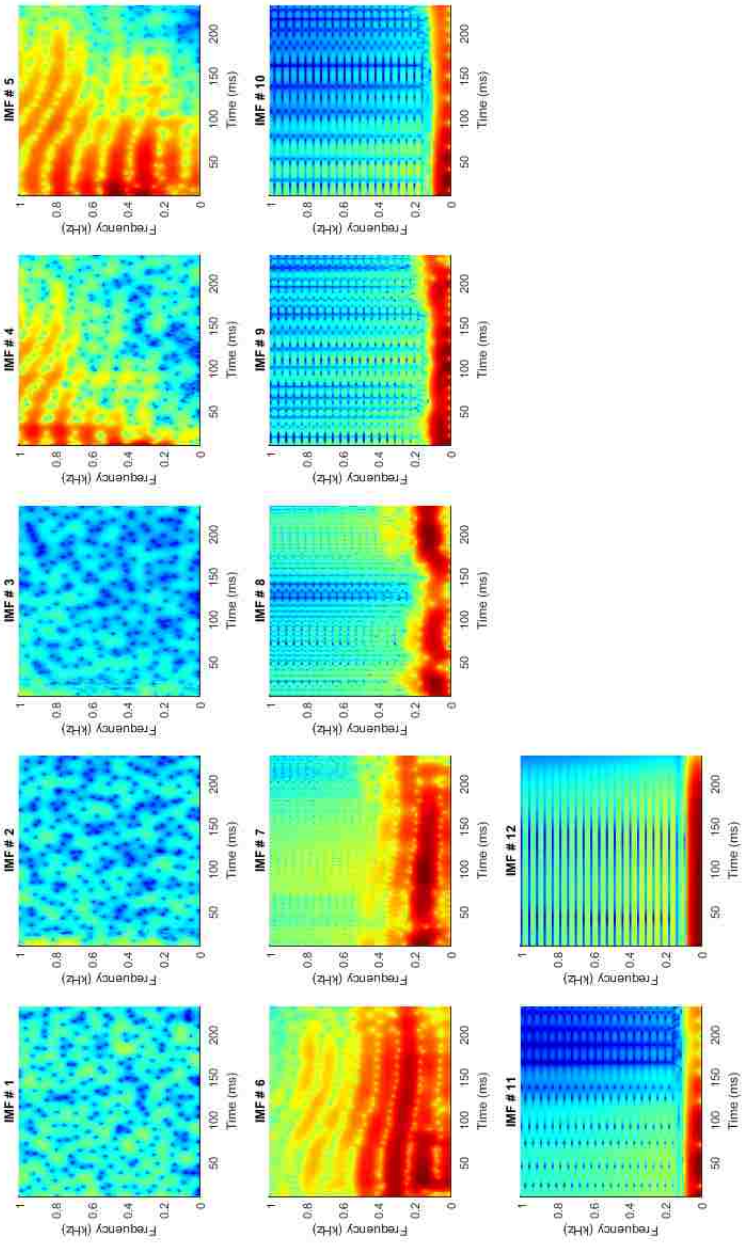


Figure A.8: Spectrograms of IMFs for signal MENRS7-feet. Range: 0 to 1 kHz

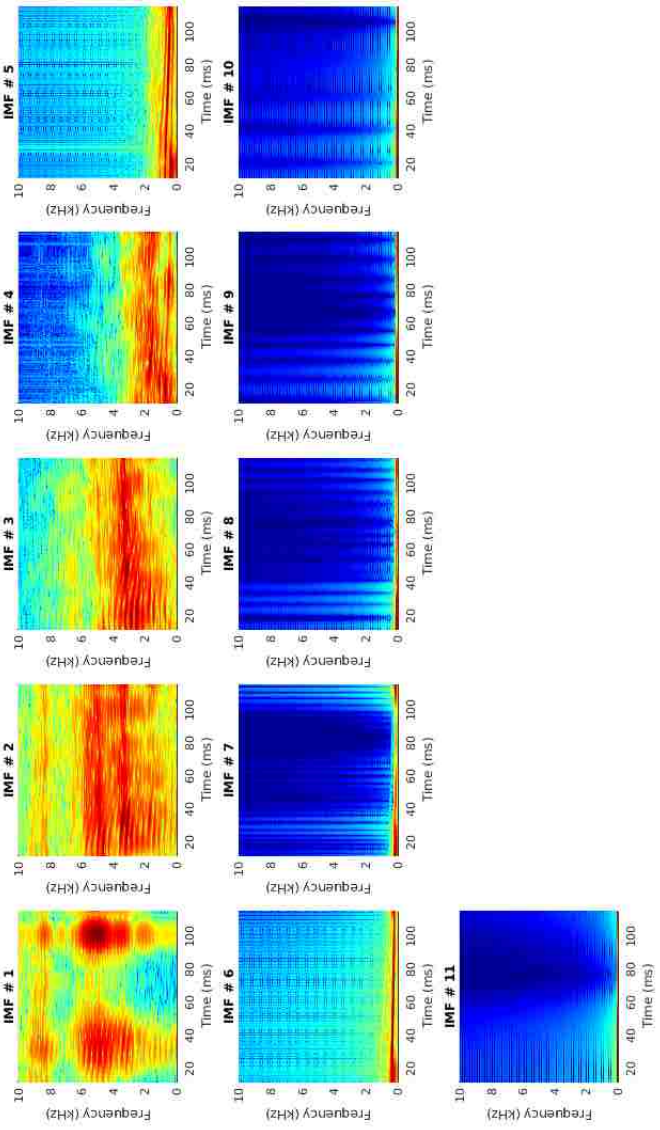


Figure A.9: Spectrograms of IMFs for signal CHILDRS1-feet. Range: 0 to 10 kHz

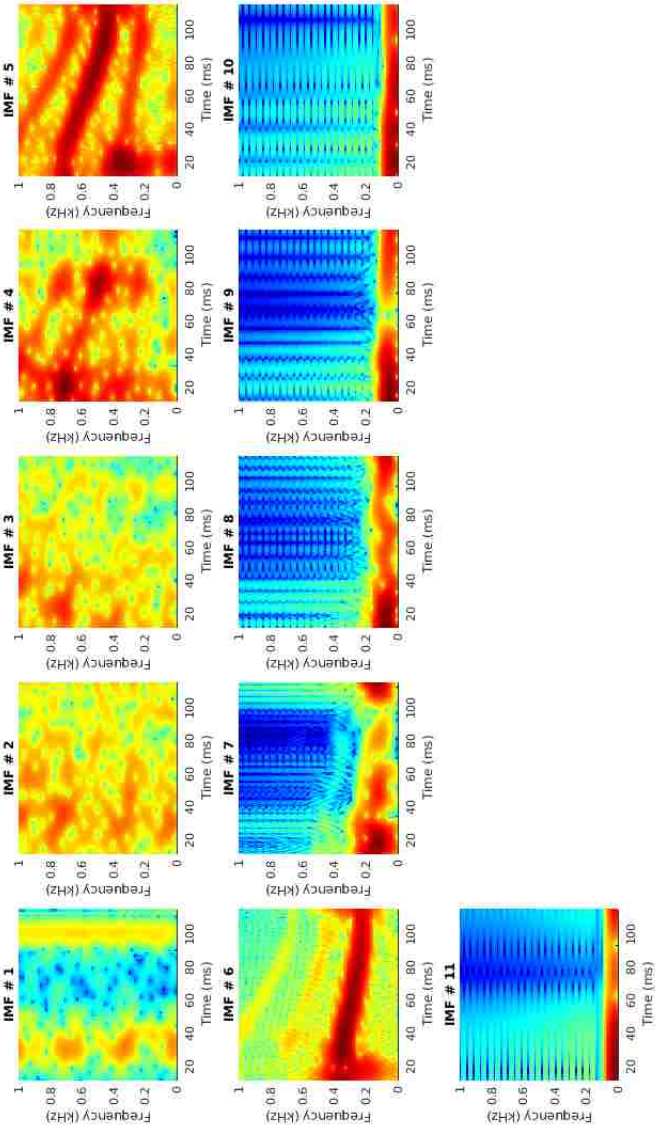


Figure A.10: Spectrograms of IMFs for signal CHILDRS1-feet. Range: 0 to 1 kHz

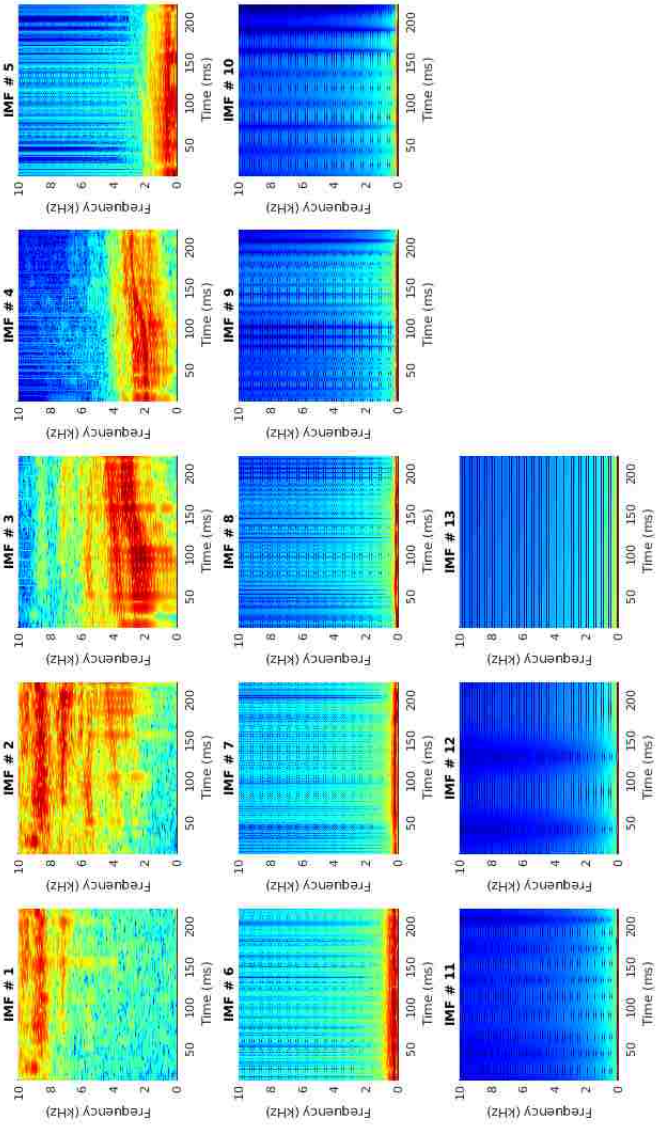


Figure A.11: Spectrograms of IMFs for signal CHILDRS6-street. Range: 0 to 10 kHz

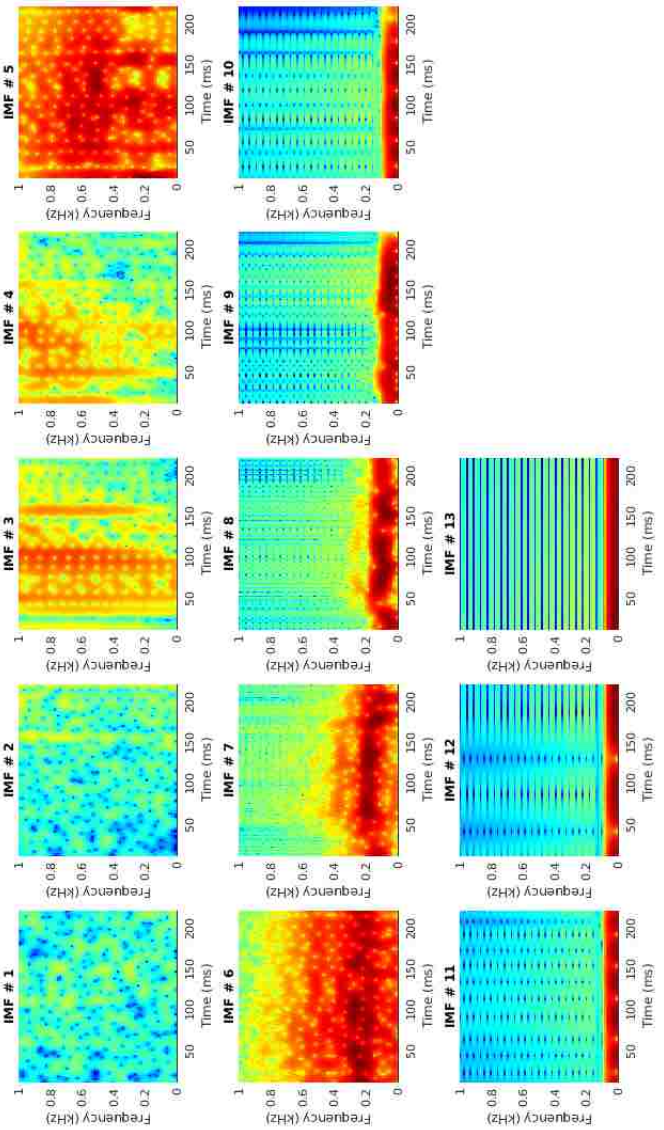


Figure A.12: Spectrograms of IMFs for signal CHILDRS6-street. Range: 0 to 1 kHz

A.3 Teager-Kaiser energies of analyzed signals

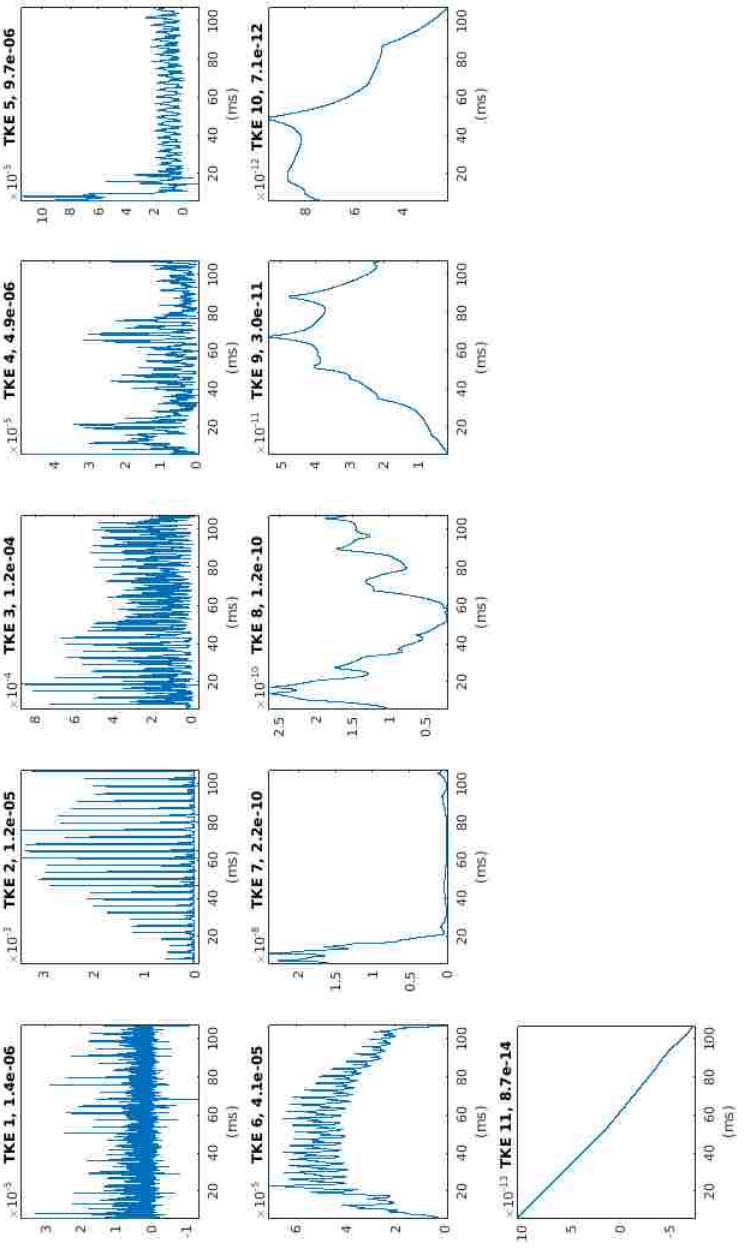


Figure A.13: Teager-Kaiser energies of IMFs for signal WOMENRS1-seeds

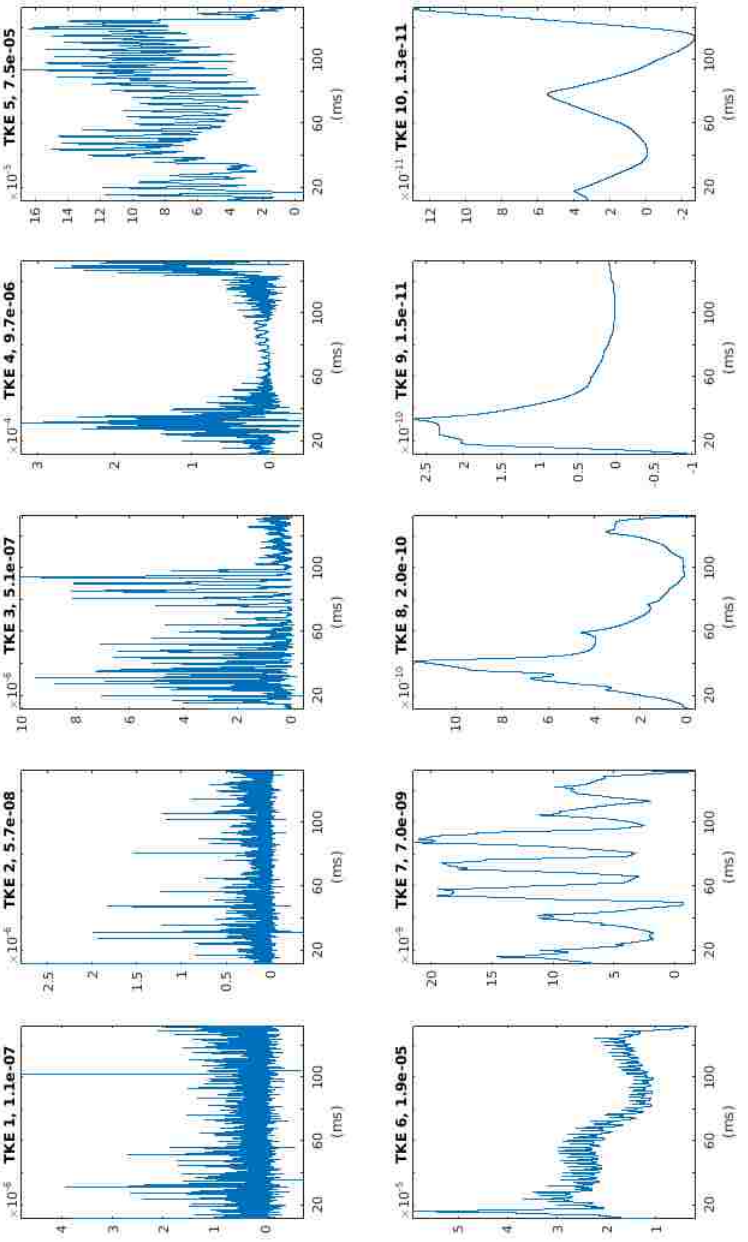


Figure A.14: Teager-Kaiser energies of IMFs for signal WOMENRS6-see

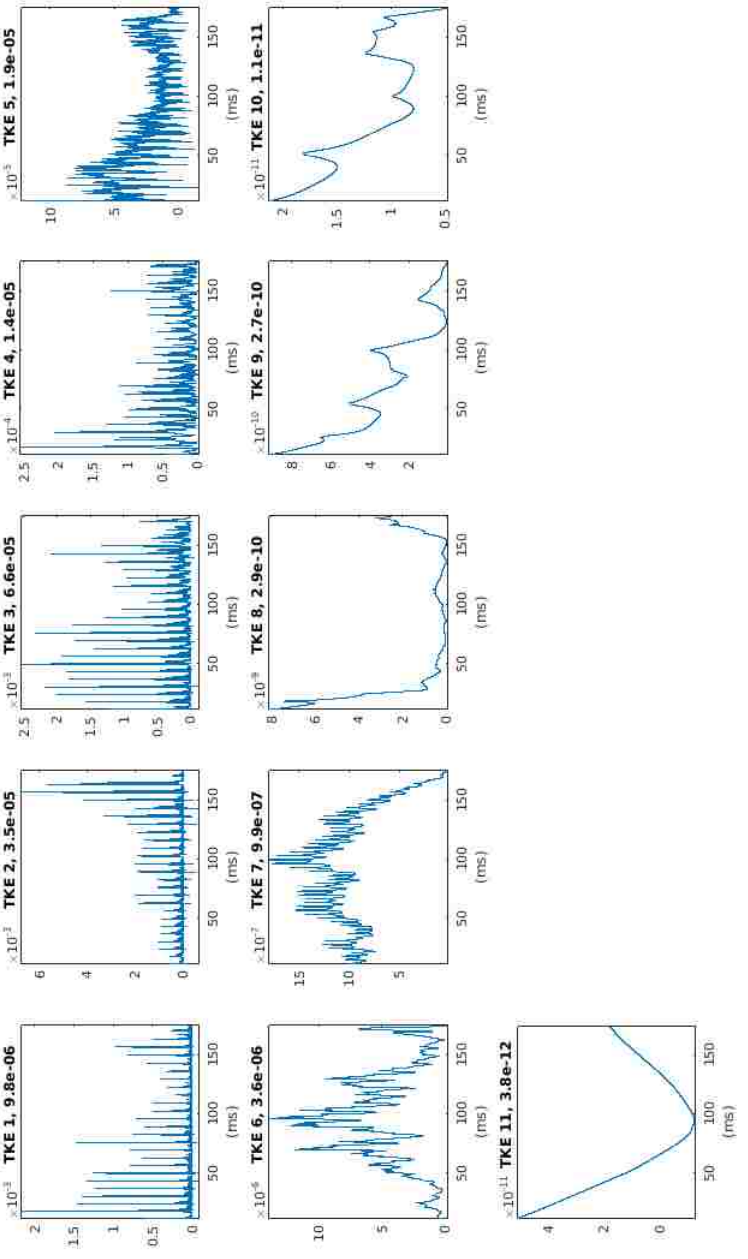


Figure A.15: Teager-Kaiser energies of IMFs for signal MENRS1-feed

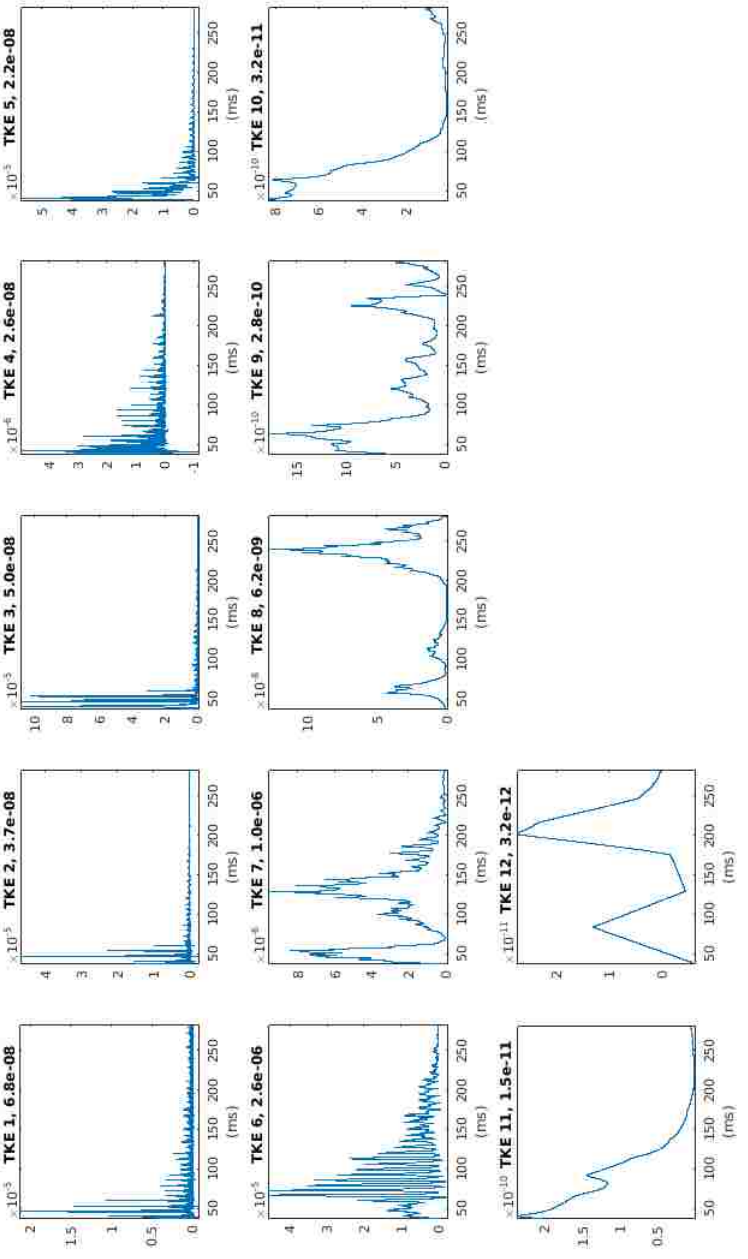


Figure A.16: Teager-Kaiser energies of IMFs for signal MENRS7-feet

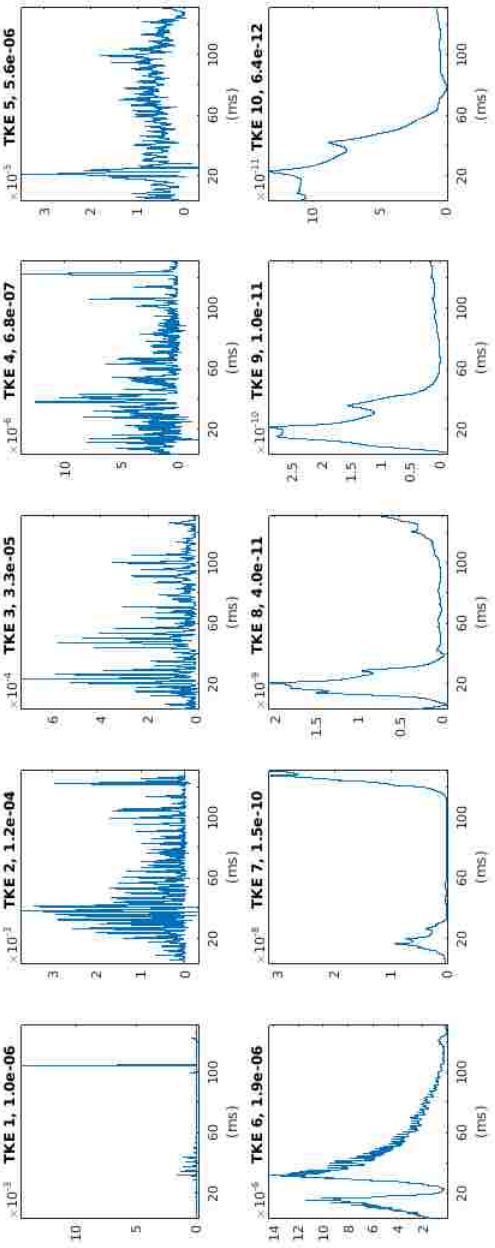


Figure A.17: Teager-Kaiser energies of IMFs for signal CHILDRS1-feet

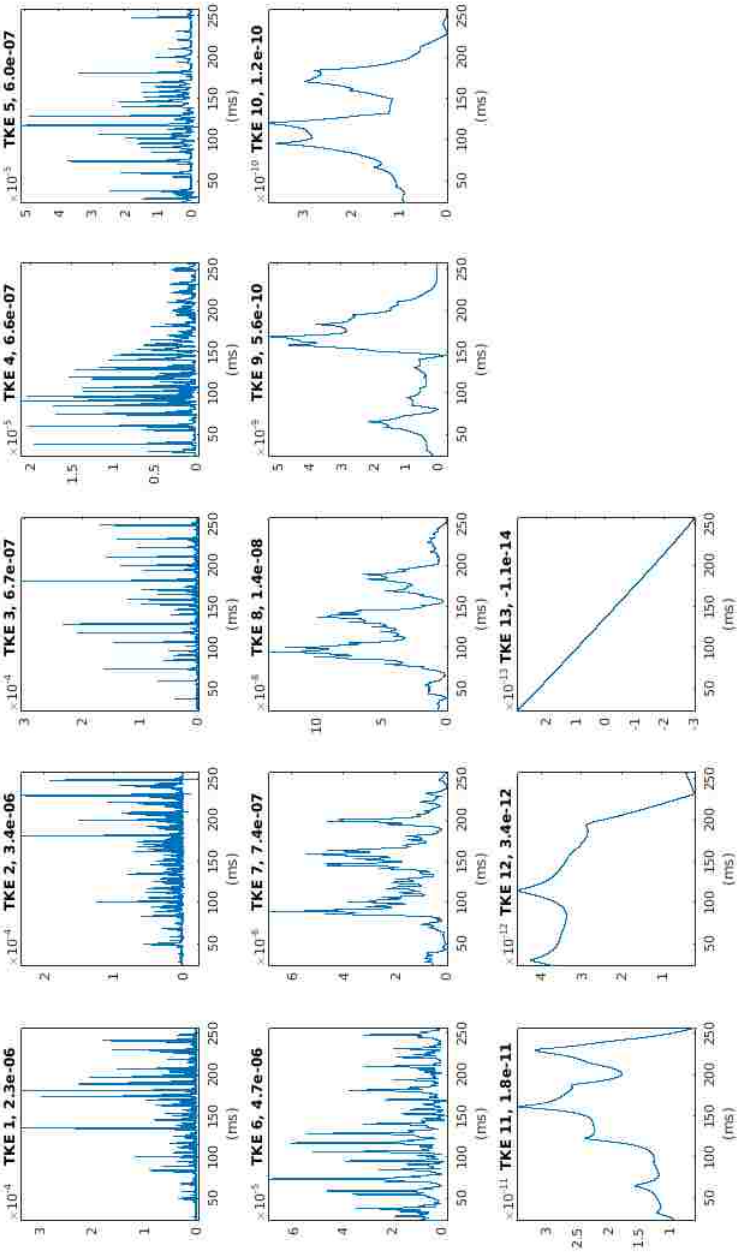


Figure A.18: Teager-Kaiser energies of IMFs for signal CHILDRS6-street

Bibliography

[1] A. Neel,"Lecture on Acoustic Phonetics 1," *Introduction to Phonetics*, SHS-303, April, 2016.

[2] L.R. Rabiner and R.W. Schafer, "Theory and Applications of Digital Speech Processing," *Pearson Higher Education Inc.*, 1042 pp., 2011.

[3] A.W. Kummer, "Cleft Palate and Craniofacial Anomalies, Effects on Speech and Resonance, 3e" *Delmar, Cengage Learning*, Clifton Park, N.Y., 768 pp., 2014.

[4] J.A. Robbins, J.A. Logeman, and H.S. Kirshner, "Swallowing and speech production in Parkinson's disease," *Annals of neurology*, 1986 Mar;19(3):283-7.

[5] Alice S-Y. Lee, Valter Ciocca, and Tara L. Whitehill, "Acoustic Correlates of Hypernasality," *Clinical Linguistics and Phonetics*, Vol. 15, No. 4-5, pp. 259-264, 2003.

[6] R. Kataoka, D. W. Warren, D. J. Zajac, R. Mayo, and R. W. Lutz, "The Relationship Between Spectral Characterisitcs and Perceived Hypernasality in Children," *Journal of the Acoustic Society of America*, Vol. 109, No. 5, Part 1, pp. 2181-2189, 2001.

[7] J. Dodderi, M. Narra, S.M. Varghese, and D.T. Deepak, "Spectral Analysis of Hypernasality in Cleft-Palate Children: A Pre-Post Surgery Comparison," *Journal of Clinical and Diagnostic Research*, Vol. 10(1):MC01-MMC02, 2016.

[8] A.P. Vogel, H.M. Ibrahim, S.Reilly, and N. Kilpatrick, "A Comparative Study of Two Acoustic Measures of Hypernasality," *Journal of Speech, Language, and Hearing Research*, Vol.52, pp. 1640 - 1651, December 2009.

[9] J.H. McClellan, C.S. Burrus, A.V. Openheim, T.W. Parks, R.W. Schafer, H.W.Schuessler, "Computer-Based Exercises for Signal Processing using Matlab 5." *Prentice Hall, N.J.*, 404 pp., 1998

[10] E. Kvedalen, *Signal processing using the Teager Energy Operator and other nonlinear operators*, Cand. Scient Thesis, University of Oslo, Dept. of Informatics, 2003.

[11] H.M. Teager and S.M. Teager, *A Phenomenological Model for Vowel Production in the Vocal Tract*, ch.3, pp. 73-109, San Diego, CA:College-Hill Press, 1983.

[12] H.M. Teager and S.M. Teager, "Evidence of Non-Linear Sound Production Mechanisms in the Vocal Tract," *Speech Production and Speech Modeling*, edited by W.J. Hardcastle and A. Marchal (Kluwer Academic, Boston, MA), pp. 241-266, 1990.

[13] T.F. Quatieri, *Discrete-Time Speech Signal Processing, Principles and Practice*, ch.9, pp. 572-3, Upper Saddle River, NJ:Prentice Hall Inc., 2002.

[14] J.F. Kaiser, "On a Simple Algorithm to Calculate the 'energy' of a Signal," *Proc. IEEE ICASSP-90*, Albuquerque, NM, pp. 381-384, Apr. 1990.

[15] A.Potamianos and P. Maragos, "Time-Frequency Distributions for Automatic Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, Vol. 9, No. 3, pp. 196 - 200, March 2001.

[16] P. Maragos, J.F. Kaiser, and T.F Quatieri, "Energy Separation in Signal Modulations with Applications to Speech Analysis," *IEEE Transactions on Signal Processing*, Vol. 41, No. 10, pp. 3024 - 3051, October 1993.

[17] A.Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *J. Acoust Soc. Am. 99 (6)*, pp. 3795 - 3806, June 1996.

[18] B.Santhanam and P. Maragos, "Energy Demodulation of Two-Component AM-FM Signal Mixtures," *IEEE Signal Processing Letters*, Vol.3, No. 11, pp. 294 - 297, November 1996.

[19] B.Santhanam and P. Maragos, "Multicomponent AM-FM Demodulation via Periodicity Based Algebraic Separation and Energy-Based Demodulation," *IEE Transactions on Communications*, Vol.48, No. 3, pp. 473 - 490, March 2000.

[20] D.A. Cairns, J.H.L Hansen, and J.F. Kaiser, "Recent advances in hypernasal speech detection using the nonlinear Teager energy operator," *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, Vol:2, pp. 780-783. IEEE.

[21] D. Gabor,"Theory of Communication," *Journal of the Institute of Electrical Engineers*,, 93, pp. 429 - 457, 1946.

[22] M.E. Torres, M.A. Colominas, G. Schlotthauer, and P. Flandrin, "A complete ensemble empirical mode decomposition with adaptive noise," *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4144 - 4147, May 2011.

[23] R.Deerling and J.F. Kaiser, "The use of a masking signal to improve empirical mode decomposition,"Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005, Vol. 4, pp. 485 - 488, March 2005.

[24] M.A. Colominas, G. Schlotthauer, and M.E. Torres, "Improved complete ensemble EMD: A suitable tool for biomedical signal processing," Biomedical Signal Processing and Control 14 (2014) 19-29, November 2014

[25] Balu Santhanam, "Generalized Energy Demodulation for Large Frequency Deviations and Wideband Signals," *IEEE Signal Processing Letters*, Vol 11, No. 3, pp. 341 - 344, March 2004.

[26] S. Sandoval, P.L. de Leon, and J.M. Liss, "Hilbert spectral analysis of vowels using intrinsic mode functions," *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 569 - 575, December 2015.

[27] M.M Saidi, P.O. Pietquin, and R. André-Obrecht, "EMD decomposition to discriminate nasal vs. oral vowels in French," SPPRA 2010, Feb 2010, Innsbruck, Austria. pp.128-132, 2010

[28] "American Cleft Palate-Craniofacial Association", http://www.acpa-cpf.org/education/educational_resources/speech_samples/

[29] D. P. Kuehn, P. B. Imrey, L. Tomes, D. L. Jones, M. M. O'Gara , E. J. Seaver, B. E. Smith , D. R. Van Demark, J. M. Wachtel, "Efficacy of Continuous Positive Airway Pressure for Treatment of Hypernasality", *Cleft Palate Craniofaial Journal* Vol. 39, pp. 267-276, 2002.

[30] N.E. Huang, Z.Shen, S.R. Long, M.C.Wu, H.H. Shih, Q. Zheng, N.Yen, C.C. Tung, H.H. Liu, "The empirical mode decomposition and the Hilbert spectrum for non-linear and non-stationary time series analysis", *Proc R. Soc. Lond.* A 454, pp. 903-995, 1998.

[31] "Introduction to the Empirical Mode Decomposition Method", www.mql5.com/en/articles/439

[32] www.phon.ox.ac.uk/jcoleman/phonation.htm

[33] www.pinterest.com/pin/208291551492101508/

[34] www.englishclub.com/pronunciation/phonemic-chart.htm

[35] http://ell.stackexchange.com/questions/260/pronunciation-of-beaches-and-bitches

[36] www.uni-bielefeld.de/lili/personen/vgramley/teaching/HTHS/acoustic_2010.html?__xsl=/unitemplate_2009_print.xsl

[37] www.studyblue.com/notes/note/n/spa-3011-exam-3/deck/14234683

[38] http://mamonu.weebly.com/formant-frequencies-table.html

[39] https://ccrma.stanford.edu/ jmccarty/formant.htm

[40] http://home.cc.umanitoba.ca/ robh/archives/arc0509.html

[41] http://apmr.matelys.com/Standards/OctaveBands.html