Heikki Rasilo

Estimation of vocal tract shape trajectory using lossy Kelly-Lochbaum model

Aalto University School of Science and Technology
Faculty of Electronics, Communications and Automation

Master's Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Technology.

Espoo, May 05, 2010

Supervisor
Prof. Unto K. Laine

Instructor
M. Sc. Okko Räsänen

Aalto-yliopisto
Teknillinen korkeakoulu

AALTO YLIOPISTO
TEKNILLINEN KORKEAKOULU

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Heikki Rasilo

Työn nimi: Ääntöväylän liikeradan arviointi käyttäen häviöllistä Kelly-Lochbaum mallia

Päivämäärä: 5.5.2010 Kieli: englanti Sivumäärä: 6+73

Elektroniikan, tietoliikenteen ja automaation tiedekunta

Signaalinkäsittelyn ja akustiikan laitos

Professuuri: Akustiikka ja äänenkäsittely Koodi: S-89

Valvoja: Prof. Unto K. Laine

Ohjaaja: DI. Okko Räsänen

On esitetty teorioita, joiden mukaan puheen ymmärtämistä helpottaa aikaisempi kokemus puheen tuottamisesta. Muuntamalla akustinen puhesignaali hypoteesiksi puhujan artikulaatioeleistä voidaan saavuttaa puhujariippumattomampi ja äänteitä paremmin erotteleva kuvaus puheesta. Tämä työ esittelee metodin, jolla ääntöväylän liikeratoja voidaan arvioida suoraan puhesignaaleista.

Tässä työssä luodaan Kelly-Lochbaum-tyyppinen ääntöväylämalli käyttäen apuna puheentuottamisen teoriaa. Malli on varustettu huulisäteilyllä ja säädettävällä huulten pituudella. Mallia käyttäen luodaan hakutaulukko, joka kuvaa vastaavuuksia puheen hetkellisten spektriominaisuuksien ja artikulatoristen muotojen välillä. Hakutaulukkoa voidaan käyttää mappaukseen akustisen ja artikulatorisen avaruuden välillä.

Luotua mallia käytetään ääntöväylän liikeratojen arvioinnissa jatkuvan puheen aikana. Liikeradat löydetään käyttämällä yksinkertaista optimointialgoritmia, joka estimoi liikeradan minimoimalla artikulaatioon kuluvaa energiaa.

Avainsanat: puhe, ääntöväylä, akustiikka, liikeradan estimointi, puheen tuotto, puheen havaitseminen, artikulatorinen kartta

AALTO UNIVERSITY
SCHOOL OF SCIENCE AND TECHNOLOGY

ABSTRACT OF THE
MASTER'S THESIS

Author: Heikki Rasilo

Title: Estimation of vocal tract shape trajectory using lossy Kelly-Lochbaum model

Date: 05/05/2010 Language: English Number of pages: 6+73

Faculty of Electronics, Communications and Automation

Department of signal processing and acoustics

Professorship: Acoustics and audio signal processing Code: S-89

Supervisor: Prof. Unto K. Laine

Instructor: M. Sc. Okko Räsänen

There are theories that during speech perception, the understanding of speech is boosted by the knowledge of the articulatory gestures based on former speech production experience. By transforming an acoustic speech signal into a hypothesis about the articulatory gestures of the speaker, it is possible to obtain a more accurate, speaker-independent description of speech. This thesis introduces a method of estimating vocal tract trajectories from speech signals.

Using the theory of speech production, a lossy Kelly-Lochbaum vocal tract model equipped with lip radiation impedance and variable lip rounding length is created. A lookup table consisting of correspondences between spectral qualities of instantaneous speech signals and articulatory shapes is created using this model. The lookup table can be used to perform acoustic-to-articulatory mapping.

The obtained model is used in estimation of vocal tract shape trajectories in continuous speech. Smooth and minimum energy trajectories are found by using a simple optimization algorithm.

Keywords: speech, vocal tract, acoustics, trajectory estimation, speech production, speech perception, acoustic-to-articulatory mapping

Preface

This master's thesis began with an aroused interest towards speech research in a course of speech processing by Unto K. Laine. I needed to start the work on my thesis as soon as possible due to the outdating degree system, and Unto soon found an interesting subject to look into from his inexhaustible source of research ideas.

First of all, I want to thank my supervisor and instructor, Unto K. Laine and Okko Räsänen. The work was done with a limited schedule, but the excellent supervision and exhilarating attitude of Unto and Okko made this work possible. I want to thank them also for a valuable lesson on writing scientific English in the form of careful revision during the last phases of this thesis, as well as encouragement and participation in writing and submitting a conference paper based on the research done during this thesis.

I also want to thank Pertti Palo for a view on on-going MRI-research. Thanks to Petri for keeping me caffeined up, and to Mark, Bruce and Scotch for keeping my spirit levels high. I want to thank Jessie for a nice mid-thesis holiday in Belgium, and my family and friends for all the support.

Espoo, 5.5.2010

Heikki Rasilo

Contents

Abbreviations and Symbols

ASR	automatic speech recognition
CV	consonant-vowel pair
DC	direct current
DCT	discrete cosine transform
F1-F2 plot	a figure illustrating the two first formant frequencies of a speech signal
FIR	finite impulse response
FFT	fast Fourier transform
IDCT	inverse discrete cosine transform
KL-model	Kelly-Lochbaum vocal tract model
LP	linear prediction
MRI	magnetic resonance imaging
MSE	mean square error
STFT	short-time Fourier transform
VC	vowel-consonant pair
VT	vocal tract
$A(x)$	cross-sectional area of the vocal tract at distance x from glottis
p	sound pressure
U	volume velocity
ρ	density
c	speed of sound
Z	acoustic impedance
T	transmission coefficient
R	reflection coefficient
E_k	kinetic energy
E_p	potential energy
f	frequency
L	inductance
C	capacitance
$\boldsymbol{T}$	a lookup table of correspondences between vocal tract shapes and formant frequencies

1 Introduction

Speech is a form of human communication, consisting of the production of patterns of a wide variety of sounds. Speech production is a delicate process where the physiological components of the *articulatory system* are changed due to muscular movements. The ability to create a massive quantity of different sounds, words and tones of voices makes speech a superior tool to pass on emotions and messages.

In today's technological applications, different forms of speech processing have received a lot of attention. Study of speech signals is needed in a wide variety of research fields, such as health care, automation, linguistics, security and mobile communications to name but a few. *Speech recognition*, *speaker recognition*, *speech coding*, *voice analysis* and *speech synthesis* are examples of subfields of speech processing, allowing the development of applications like voice controlled systems, text-to-speech and speech-to-text transformation, applications for disabled people, user identification based on voice, and speech signal compression.

Speech recognition systems have originally been based purely on the acoustic characteristics of speech signals. However, a lot of research is being carried out today in the field of speech concerning *multimodality*. The term refers to the observation that different kinds of stimuli, for example visual and auditory characteristics can be used to increase the reliability of systems. For example visible articulatory movements have been used in addition to the audio signal to support *automatic speech recognition, ASR* [1].

Speech is a skill learnt usually in early childhood. As a whole, it still remains a mystery how a small child can manage to learn a whole language in such a perfection during a small amount of time. Nevertheless, research has shown that in addition to hearing the people around speak, it is important to actually see how speech is performed. During the first six months of life, infants try to imitate the *articulatory gestures* of people speaking, thus attempting to learn the patterns of acoustic characteristics of speech sounds and the corresponding articulatory movements [2]. The importance of visual perception of speech can be also noticed when for example in an animation the lip movements do not correspond to the spoken sound. This may disturb the understanding of speech.

In this thesis, the dependencies between speech perception and production are examined. The main question in focus is whether it is possible to obtain useful information about articulatory movements based on analysis of continuous speech signals. If this would be possible, according to the motoric theory of speech production, the information regarding the dynamics of the vocal tract could be used to reduce the uncertainty in the automatic classification of speech related patterns such as speech sounds and words.

Exact information about the dynamic vocal tract could be a step towards *speaker normalization,* allowing mapping of acoustically differing speakers in the same universal space. Vocal tract shape tracking would also give new tools for speech recognition in a situation where multiple people are speaking simultaneously. The acoustical characteristics of different speakers may be very similar, but managing to track the continuous vocal tract movements for each person might identify the speech

performed by them. This effect has also a link to real human communication, where the understanding of speech of a certain person in a crowded environment is boosted by looking directly at the person.

In this thesis, vocal tract profiles are estimated from speech by creating an extensive *lookup table* that provides information about the hypothetical vocal tract shapes corresponding to a sound with certain spectral characteristics. The fundamental property of this *acoustic-to-articulatory mapping* is that it is not a unique, *one-to-one,* mapping but leads to a non-unique, *one-to-many,* problem indicating that one spectral shape can be created by many different vocal tract configurations.

The lookup table is created by varying the vocal tract shapes around the *anchor shapes* found for eight Finnish vowels. The anchor shapes are obtained by starting from existing accurate vocal tract area functions for the English vowel sounds and manually reshaping them to better represent the Finnish vowel sounds from acoustical, spectral and articulatory point of view. The English speech sounds are used as a starting point since accurate measured vocal tract shapes for Finnish vowel sounds do not yet exist.

Spectral representations corresponding to different vocal tract configurations are obtained by using a Kelly-Lochbaum model. A Mathematica implementation of a KL-model by Unto K. Laine was used as a starting point. The program has been in use in some European universities for speech research purposes. For the sake of simplicity and low computational complexity, the amount of parameters is kept at a minimal level that still enables sufficiently reasonable synthesis of Finnish vowels. Viscous, thermal or wall vibration losses are not taken into account but the most important loss, namely *lip radiation impedance*, is implemented, making the vocal tract model a more realistic, *lossy,* model. Also the length of the section at the lips can be varied. In total, the vocal tract model uses 17 parameters: 16 uniform sections to define the vocal tract area function and one parameter for the lip rounding.

Variation in the anchor shapes is created in a physiologically reasonable manner in order to form formant clusters around the vowel anchor points in the formant domain. The lookup table itself gives interesting information about the characteristics of vowel sounds. The obtained formant space contains a clear "*vowel triangle*" that seems to be divided further into two main clusters of back and front vowels. General information regarding where the formant frequencies of each anchor shape are likely to move is also obtained in the process.

Tracking of the dynamic area function starts with extracting the changing formant frequencies of the continuous speech sample in question. A certain number of candidates for the actual vocal tract shape at each time window is then selected based on the estimated formant frequencies. A smooth, minimum effort path through the articulatory space is then searched by selecting those candidates that best fit the given criteria (smoothness and minimum effort).

The basic idea of this study was to start with the anchor shapes of the eight Finnish vowels and create articulatory variability around them in order to construct a look-up table with the corresponding formant frequencies. The table is then used to create hypotheses of the possible articulatory state and movements from the formant frequencies of the continuous speech under analysis. The hope was that this approach could provide important information not only about vowel like sounds but even related to the *consonant-vowel, CV*, and *vowel consonant, VC*, transitions. Thus the method could reveal important additional information associated with consonants. E.g., the

place of articulation is an important cue when clustering and characterizing consonant sounds. Even vowel sounds of continuous speech tend to differ from their variants pronounced in isolation due to the energy minimization and conserving property of articulatory movements. Thus the method could help to study, model and understand also the variability of vowels in different articulatory contexts.

The thesis is constructed as follows: In chapter 2 some of the history and the former research related to the topic are shortly introduced. In chapter 3 the theoretical backgrounds on the methods used in this work are discussed and chapter 4 describes the construction of the vocal tract model based on the theoretical background. In chapter 5 the creation of the lookup table is discussed and in chapter 6 the vocal tract shape tracking methods are explained. Chapter 6 also reports tests performed in the task of tracking of vocal tract shapes of continuous speech. Finally, conclusions are drawn and challenges for future work are outlined.

2 A view on former research

The possibility to create speech-like sounds artificially has been in focus of research for a few centuries. A challenge of building a machine being able to reproduce spoken vowels and explain their physical properties was issued by the Royal Academy of Sciences in St. Petersburg in 1779. In response to this, Wolfgang von Kempelen was able to construct a machine consisting of a wooden box connected to bellows serving as *lungs* on one side, and to a rubber funnel serving as a *vocal tract* on the other. A vibrating reed before the vocal tract created the *glottis excitation*. A well-trained user could produce astoundingly realistic vowel sounds at that time by controlling the funnel with one hand, and the valves in the wooden box with the other. [3]

In the 1930's, attempts to generate an electrical synthesizer for speech were made. One of the most important ones was *Voder*, voice operation demonstrator. The device used several electrical circuits to create resonances at certain frequencies. These bandpass filters were spread across the important speech frequency range and their gains could be individually controlled by fingers. The excitation was created by a noise source for unvoiced sounds or by a relaxation oscillator for voiced sounds. This machine was demonstrated at the World's Fairs of 1939 in New York. [4], [5]

From Voder, the development went forwards towards a device that could derive important parameters from analyzed speech signal, and the signal could then be recreated by using these parameters. Such a method was introduced by Homer Dudley in 1939, and the device under development was given the name Vocoder, Voice coder. From the speech, the *fundamental frequency*, and the spectrum were analyzed. Different versions of the Vocoder used different amounts of spectrum analyzing channels. Each channel consisted of a bandpass filter and a rectifier to measure the signal power in the band. In the synthesis phase, the gathered information was fed to a synthesis filter bank, and a similar to the original speech-signal was obtained. This device was the beginning of speech compression, using only a limited amount of parameters to capture the most important characteristics of speech signals. [6]

One of the first attempts to create an articulatory system to mimic human speech production was introduced by Flanagan, Ishizaka and Shipley in 1979. Their model consisted of several parameters allowing the control of larynx and vocal tract shape coupled with nasal tract, having a direct association with the physiology of human speech production. The vocal tract itself consisted of ten cross-sectional tubes, where the larynx tube area was fixed, and lip rounding and larynx tube lengths were equal to one tenth of the vocal tract length. The mouth area was manually observed from the speaker. In their work, a recorded speech signal was analyzed and the spectral difference between the speech and the synthetic signal obtained by the articulatory model was minimized by adjusting the parameters of the model. This type of approach for estimating articulatory parameters is nowadays referred to as *analysis-by-synthesis.* [7], [8]

The parameters of the optimization loop used in analysis-by-synthesis techniques require good initial parameters close to the global optimum. From this starting point, the optimization loop finds the local minimum of the given *cost function* that is being minimized. Selection of the initial parameter values can be done using *articulatory codebooks* or *lookup tables*, i.e., mappings between acoustic and articulatory

parameters. This method was used in several articulatory synthesizers, one of them being the articulatory speech mimic by J. Schroeter, J. N. Larar and M. M. Sondhi, presented in 1987 [9]. The mimic tried to associate the vocal tract and glottal model parameters by a two-stage procedure. Vocal tract model was estimated first keeping glottal parameters fixed, and then glottal parameters were estimated keeping the vocal tract fixed. The articulatory codebook was created using Mermelstein's articulatory model [10], transforming vocal tract characteristics into an *area function* in order to describe cross-sectional areas from glottis towards the lips. All the vowels were modeled by matching formant and vocal tract shape data. Consonants and nasals were modeled using related physiological features. In total, 20 vocal tract shapes were used as root shapes, from which interpolation was used to obtain a total of 10090 shapes. The interpolation was done in straight lines in parameter space from one root shape to another. [9], [11]

Creation of articulatory codebooks has proved to be a challenging task, and researchers have studied numerous different methods to get reliable and physiologically realistic mappings between speech signals and vocal tract shapes with a certain glottal excitation. The main problem with using codebooks is that the acoustic-to-articulatory mappings are not unique, meaning that several different articulatory parameters may lead to a single vocal tract transfer function. [8]

Fundamental work on articulatory-to-acoustic mapping has been done by Atal *et al.* in the 1970's [12]. They used computer sorting to obtain vector pairs of articulatory parameters and formant frequencies. A lossy vocal tract model was used and the transfer function was calculated numerically with a computer. In Atal's work, so-called *fibers* were described, referring to articulatory regions that are mapped into a single point in the acoustic space. The area function of the vocal tract was controlled by five parameters: the distance of the maximum constriction from the glottis, the cross-sectional area of the maximum constriction, the area of the lip opening, lip rounding length, and the vocal tract length.

There are several possibilities for obtaining geometric vocal-tract data for mappings. For example X-ray and ultrasound measurements and magnetic resonance imaging, MRI, have been used for these purposes [13]. Some methods expose the subject of the measurements to a health risk and in some cases it is impossible to record accurate speech signal simultaneously because of the magnetic interference or acoustic noise. Today, mainly MRI-methods have been in focus of research, because they give accurate measurements for the physiological structure of the vocal tract, and the drawbacks due to strong resonating magnetic fields and noise are likely to be overcome by carefully designed equipment. [14]

Research for measuring speech during MRI scan of the vocal tract is done currently in the Department of mathematics and system analysis in Aalto University School of Science and Technology by J. Malinen and P. Palo. A recording arrangement is being constructed allowing accurate measurements despite the noise and magnetic field inside the MRI-machine. The results can be used further in numerical models for the vocal tract. [14]

Most tools for articulatory tracking have been unable to maintain the two main criteria needed based on physiological speech production: the acoustical proximity of the vocal tract shapes to the original data, and the smoothness of articulatory movements [15]. However, new tools for optimization and calculus can be exploited to

combine the two criteria as can be noticed from the recent studies. Laprie and Mathieu [15] have approached the problem using Maeda's articulatory model [16] and improved lookup algorithm for the created codebooks. Three articulatory parameters, jaw position, tongue dorsum position and tongue shape, were tracked in vowel transitions, such as /iai/ and /iui/. Dang and Honda [17] have estimated vocal tract shapes using a 3D physiological articulatory model. The model is equipped with physiological muscular configuration and thus takes into account the physiological constraints in human articulation.

Carré has carried out research to find explanations for the phonologies of the world's languages from the evolutionary point of view. Carré has examined speech production characteristics by introducing changes to a uniform acoustic tube that is 18 cm long [18]. The formant frequencies according to the vocal tract shapes are calculated by an algorithm proposed by Badin and Fant [19]. The variations in the tube shape are performed so as to produce maximal acoustic change by introducing minimal deformations. Deformations resulting to changes in the first and second formant frequencies are made, and the characteristics of the formant shifts are discussed. Carré shows that using simple acoustic principles, the basic characteristics of the speech production system can be derived.

In 2009, Carré has created an acoustic space by further varying the 18 cm long tube [20]. From the uniform configuration, iterative deformations to the vocal tract shape are made considering the first two formants, and formant frequency trajectories are seen to form the well-known vowel triangle. Since the vowels mainly used in languages occur at the borders of this triangle, i.e., acoustically far from each other, it seems likely that the formation of vowel sounds has evolved for communication purposes. Carré states that "*Each point in the vowel triangle can, in fact, be reached by the articulatory machinery, but the transition from one vowel to another, i.e., the dynamics, results from strategies favoured by geometrical and acoustical properties of the tube*".

The mentioned dynamic properties of the articulatory system are further examined in this thesis. The formation of the vowel triangle is approached in a slightly different manner than in Carré's study. Deformations are introduced to the vocal tract shapes of known Finnish vowel sounds. Formant frequencies are estimated by using a Kelly-Lochbaum model terminated with lip radiation impedance. Additional parameter to Carré's model is the varying length of lip rounding, which adds important information about the formation of certain vowel sounds. Deformations corresponding to four formant frequencies are made. The objective is to be able to track vocal tract shape transitions from recorded speech signals. An optimization algorithm is implemented to estimate the dynamics of the vocal tract movements.

3 Theoretical background for vocal tract models

In this chapter, the physiology of the human vocal tract and the important physical characteristics of sound propagation are discussed. This chapter emphasizes the theoretical background that has to be adopted in order to understand the methods that are widely used in speech research and vocal tract models. These observations and results will be used later when creation of a lossy Kelly-Lochbaum model of a vocal tract is explained in chapter 4.

3.1 Speech production mechanism and speech perception

Human speech is produced by a delicate system consisting of respiratory organs, including lungs and the trachea, and the vocal tract, including larynx, mouth and nasal tract. In addition, a plethora of small organs exist that affect the sound production throughout the vocal apparatus. Here, a short description of the main aspects affecting speech sound production is given.

During expiration, the diaphragm causes the air pressure in the lungs to increase and forces the air to flow towards the vocal tract. At the larynx, the vocal cords, whose tenseness can be controlled by muscular movement, can be set up to create quasi-periodic pulses of air that move into the vocal tract to create voiced sounds. These pulses are formed by the pressure building up below the closed vocal cords, and its fast reduction when the pressure is high enough to fold the vocal cords apart. Due to Bernoulli's principle, increment in the flow speed of a fluid occurs simultaneously with the decrease in pressure, causing the vocal cords to return to their original position and the pressure to start building up again. The space between the vocal folds is called *glottis.* The most important components of the speech production system can be seen in Figure 1.

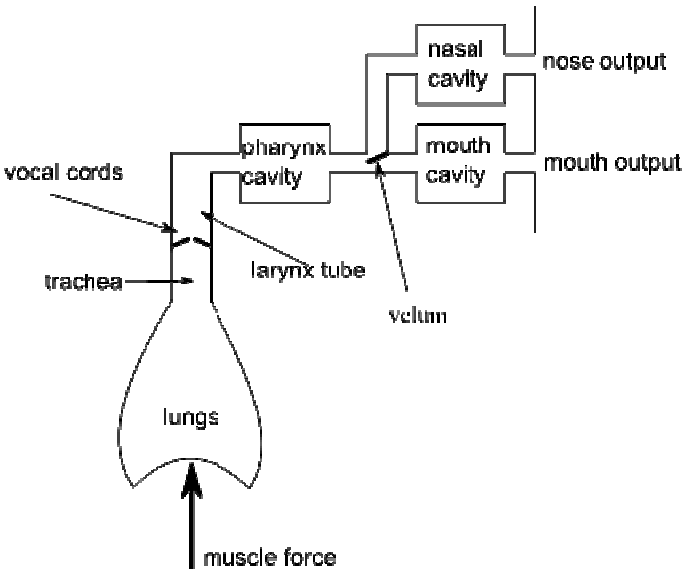


Figure 1. Components of the human speech production mechanism as illustrated by Flanagan. [19]

The quality of the voiced sounds depends mainly on the shape of the articulatory system situated after the larynx. Control of the muscles of the articulatory system has to be learnt before the continuous speech can be produced. This skill is normally developed in childhood by listening and watching people speaking, and by trying to mimic the perceived sounds. Hearing is an essential factor giving acoustical feedback for the produced sounds and helping to further develop the control of the system. [21]

All spoken languages consist of a variety of distinguishable speech sounds that can be combined to produce messages with different meanings. The basic linguistic element is called a *phoneme*, and is related to the articulatory gestures of the language. Phonemes are the smallest units in a spoken language that have a distinctive function in an utterance of the language in question. If one phoneme in an utterance is changed to another, the meaning of the utterance is changed. Phonemes are often classified according to their place and method of production in the articulatory system. Phonemes are also usually divided into two main classes: *vowels* and *consonants*.

Vowels are generally produced by voiced vocal cord excitation at the larynx. During distinct vowel sounds, the vocal tract has usually a stable position. The nasal cavity affects vowel sounds if it is coupled to the oral cavity. The coupling is controlled by the *velum*. When the velum is open during a vowel it causes *nasalization*. In the case of nasal consonants, such as /n/ and /m/, the oral cavity is completely closed and the sound propagates only through the nasal cavity.

Vowel sounds are often classified into groups of *front vowels*, *back vowels* and *central vowels* according to the position of the highest constriction in the vocal tract. The degree of the highest constriction further classifies the vowels into groups of openness. Vowel sound classification for Finnish vowels is shown in Table 1 [22].

Table 1. Classification of Finnish vowel sounds

Degree of constriction	**Position of highest constriction**		
	Front	**Central**	**Back**
High	/i/	/y/	/u/
Medium	/e/	/oe/	/o/
Low	/ae/		/a/

In speech analysis, vowel sounds are often identified by their resonance frequencies caused by the shape of the vocal tract. These resonance frequencies are called *formants.* The two first formants define the main characteristics of the vowel, and the vowel can usually be identified by using them only. The higher formants affect mostly the timbre of the vowel sound. Formants 3 and 4 are of importance mainly in front vowels. In general, three to five first formant frequencies are used when qualities of speech sounds are analyzed. [23]

In addition to vowel sounds, languages consist of less voiced sounds called consonants. Consonants are classified into a number of subgroups according to their articulatory characteristics. *Fricative consonants* are created by a turbulent air flow in a narrow constriction at some part of the articulatory system. Examples of English fricatives are the phonemes /v/ and /s/. *Stop consonants* have a complete closure at some location of the vocal tract which is then rapidly opened, causing a sudden release of air pressure. This happens for example in phonemes /p/ and /k/. *Nasal consonants*, such as /m/ and /n/ are normally excited by the vocal cords, but the oral cavity is closed at some

location. Open velum causes the air flow to be transferred through the nasal cavity. *Glides and semivowels* is a small group of consonants that resemble vowels but the vocal tract is heavily constricted. Phoneme /l/ is an example of a semivowel and /j/ an example of a glide. [21]

The term *source-filter-model* refers to the fact that the speech production mechanism can be represented as the combination of a voice source and a vocal tract filter that shapes the source signal into speech at the output. The voice source can be a voiced signal, noise signal, combination of the two, or just complete silence.

3.2 Wave propagation in the vocal tract

Modeling of sound waves in the vocal tract in an exact manner is a difficult task due to the complexity of the shape of the tract and complicated acoustical phenomena involved, such as frequency dependent losses, yielding walls, flow dynamics, jet formation and turbulence. In reality, the oral cavity and nasal cavity are lossy tubes with a continuously varying cross-sectional area, and the wave motion is almost impossible to describe accurately. With a number of simplifications of the characteristics of the vocal tract, models that are easier to handle can be created. The propagation of sound waves is discussed here in a physical sense and important solutions to be used in the further model are derived.

3.2.1 Simplification of the acoustic properties of the vocal tract and introduction to Kelly-Lochbaum model

The most important information regarding intelligibility of speech lies in the frequency range from 0 to 4000 Hz. The wave propagation calculations can be simplified by assuming that the vocal tract[1] can be straightened out and presented as a straight tube with a varying cross-sectional area. M. M. Sondhi has shown that the bending of the tract has no much effect in the frequency band of less than 4000 Hz [24]. If the diameter of the widest cross-sectional area in the tract is less than a wavelength, the propagation of higher modes is prevented and one-dimensional plane wave propagation can be assumed. In human vocal tract this assumption can be made up to 4000 Hz in frequency. Also, if the vocal tract walls are assumed to be non-yielding with no viscous or thermal losses, and conservation of momentum and mass are taken into consideration, the wave propagation can be represented with acoustic field equations

$$A(x)\frac{\partial p(x,t)}{\partial x} = -\rho\frac{\partial U(x,t)}{\partial t} \tag{1}$$

[1] In this work, the model to be created does not take the nasal cavity into account. Thus, the term vocal tract is used to describe one tube of varying area, including the larynx and the oral cavity.

$$\frac{\partial U(x,t)}{\partial x} = -\frac{A(x)}{\rho c^2}\frac{\partial p(x,t)}{\partial t} \tag{2}$$

where the cross-sectional area A is only dependent on the distance from glottis x. p and U are the sound pressure and volume velocity respectively, both depending on the location and time t. c is the velocity of sound and ρ the density of air. [8], [21], [25]

Combining these two equations leads to the widely used Webster equation

$$\frac{1}{A(x)}\frac{\partial}{\partial x}\left[A(x)\frac{\partial p(x,t)}{\partial x}\right] = \frac{1}{c^2}\frac{\partial^2 p(x,t)}{\partial t^2} \tag{3}$$

This equation can be integrated only numerically, so further simplifications are often made to shape the model more suitable for practical purposes [21]. It can be assumed that the vocal tract is constructed of a certain number of co-axial uniform cylindrical sections. This model was suggested by Kelly and Lochbaum in 1962 [26]. In this case, the parameter $A(x)$ in equations (1) and (2) is constant A within each tube section, and the equations combined simplify to following forms for volume velocity and pressure

$$\frac{\partial^2 U(x,t)}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 U(x,t)}{\partial t^2} \tag{4}$$

$$\frac{\partial^2 p(x,t)}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 p(x,t)}{\partial t^2} \tag{5}$$

The general solution for these homogenous wave equations is the sum of left and right travelling wave components. Right and left travelling pressures are referred to as p^+ and p^-, and corresponding volume velocities are U^+ and U^- respectively. Also, the dependence of both time and location can be changed to dependence only on time by using the speed of sound in transforming the location into time domain. The solution becomes

$$U(x,t) = U^+\left(t + \frac{x}{c}\right) + U^-\left(t - \frac{x}{c}\right) \tag{6}$$

$$p(x,t) = p^+\left(t + \frac{x}{c}\right) + p^-\left(t - \frac{x}{c}\right) \tag{7}$$

The travelling wave components to each direction have to satisfy the wave equation (1) separately, yielding

$$A\frac{\partial p^+\left(t + \frac{x}{c}\right)}{\partial x} = -\rho\frac{\partial U^+\left(t + \frac{x}{c}\right)}{\partial t} \tag{8}$$

Using chain rule $\frac{\partial U}{\partial t} = \frac{\partial U}{\partial x}\frac{\partial x}{\partial t}$ and the relation between location and time yields

$$\frac{\partial p^+\left(t+\frac{x}{c}\right)}{\partial x} = -\frac{\rho c}{A}\frac{\partial U^+\left(t+\frac{x}{c}\right)}{\partial x} \tag{9}$$

After integration

$$p^+\left(t+\frac{x}{c}\right) = -\frac{\rho c}{A}U^+\left(t+\frac{x}{c}\right) \tag{10}$$

Using similar steps for the left-travelling wave, and noticing that now relation $t = -xc$ has to be used because of the different direction of propagation, the pressure becomes

$$p^-\left(t-\frac{x}{c}\right) = \frac{\rho c}{A}U^-\left(t-\frac{x}{c}\right) \tag{11}$$

Acoustic impedance is defined as

$$Z = \frac{\rho c}{A} \tag{12}$$

and gives with these choices of directions the compact formulas

$$p^+ = -ZU^+ \tag{13}$$

$$p^- = ZU^- \tag{14}$$

In Figure 2, an example of vocal tract shape of a lossless Kelly-Lochbaum model is shown, as well as one junction where the pressure waves p_{1+} and p_{2+} are travelling to the right and p_{1-} and p_{2-} to the left.

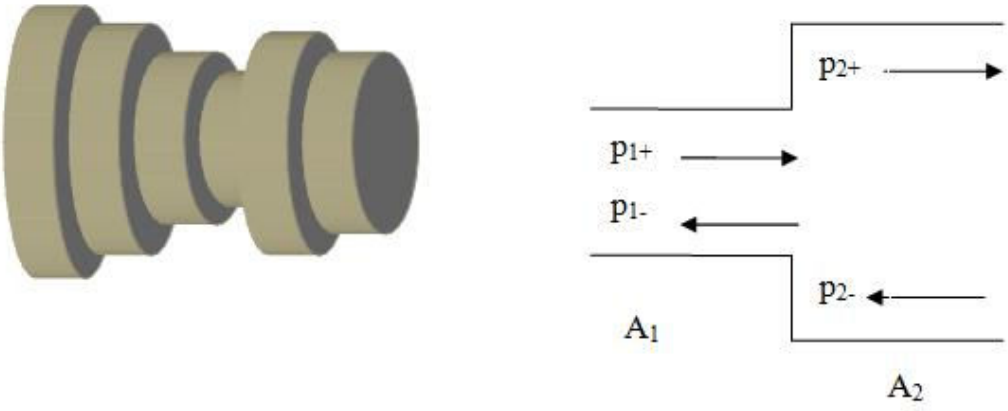


Figure 2. Approximation of the vocal tract with uniform cylindrical tubes (left), and the pressure wave propagation in one junction of the model (right).

3.2.2 Derivation of scattering equations[2] for tube junctions

The continuity conditions at a junction between two tubes are the continuity of both volume velocity and pressure through the junction

$$\begin{cases} p_1 = p_2 \\ U_1 = U_2 \end{cases} \tag{15}$$

When the pressure wave propagating right meets the junction, it is partially reflected back and partially continued to the next tube. The equations get a form

$$\begin{cases} p_1^+ + p_1^- = p_2^+ \\ U_1^+ + U_1^- = U_2^+ \end{cases} \tag{16}$$

The use of equations (13) and (14) and the introduction of reflection and transmission coefficients for pressure, *R* and *T* respectively, yield

$$\begin{cases} p_1^+ + p_1^- = p_2^+ \\ -\dfrac{p_1^+}{Z_1} + \dfrac{p_1^-}{Z_1} = -\dfrac{p_2^+}{Z_2} \end{cases} \Rightarrow \begin{cases} p_1^+ + Rp_1^+ = Tp_1^+ \\ \dfrac{p_1^+}{Z_1} - \dfrac{Rp_1^+}{Z_1} = \dfrac{Tp_1^+}{Z_2} \end{cases} \tag{17}$$

Combining these and equation (12), the reflection and transmission coefficients become

$$\begin{aligned} R &= \frac{Z_2 - Z_1}{Z_2 + Z_1} = \frac{A_1 - A_2}{A_1 + A_2} \\ T &= \frac{2Z_2}{Z_2 + Z_1} = \frac{2A_1}{A_1 + A_2} \end{aligned} \tag{18}$$

With similar calculations it can be shown that the reflection coefficient for the left travelling pressure wave is the inverse of the reflection coefficient for the right travelling wave. Thus, the junction of the Kelly-Lochbaum model can be interpreted as seen in Figure 3.

[2] In its original sense the term *scattering* usually refers to the deviation of propagating waves from their straight trajectory due to local non-uniformities in the medium. Scattering in this sense can be taken into account in more complicated vocal tract models, where wave propagation is modeled in more than one dimension. In this work, the term *scattering equation* is used to describe the equation used to calculate the one-dimensional reflection and transmission of sound waves.

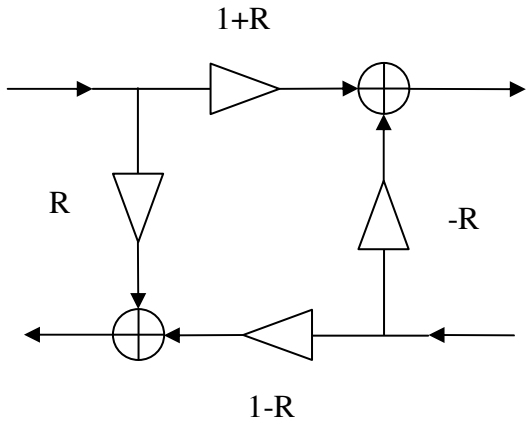


Figure 3. One junction of the Kelly-Lochbaum vocal tract model for pressure waves.

It has to be noted that this junction has been calculated for the pressure wave. The glottal excitation is typically implemented as a volume velocity source. Therefore it is desirable to derive the scattering equations for volume velocity waves. In this case it is important to notice that when volume velocity is reflected, its direction is reversed, causing $U_1^- = -RU_1^+$. Using this notation in equations (12), (13), (14) and (16) yields

$$\begin{cases} -Z_1U_1^+ + Z_1U_1^- = -Z_2U_2^+ \\ \qquad U_1^+ + U_1^- = U_2^+ \end{cases} \Rightarrow \begin{cases} -Z_1U_1^+ - Z_1RU_1^+ = -Z_2TU_1^+ \\ \qquad U_1^+ - RU_1^+ = TU_1^+ \end{cases} \tag{19}$$

Because the reflection coefficient from left to right is maintained as R, the bottom equation shows that the transmission coefficient will now be $T = 1 - R$ from left to right. The KL-junction for volume velocity will thus be as illustrated in Figure 4.

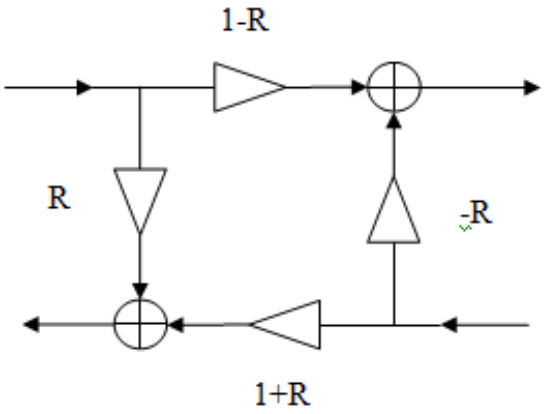


Figure 4. One junction of the Kelly-Lochbaum vocal tract model for volume velocity waves.

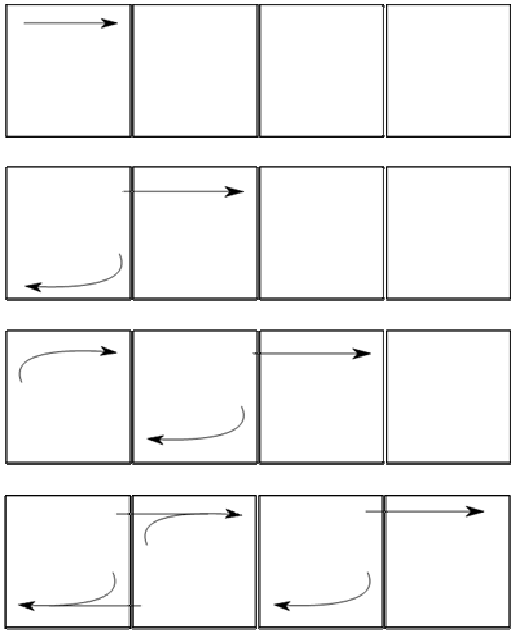
3.2.3 Choice of sampling frequency

The sampling frequency for the Kelly-Lochbaum model has to be chosen to match the total delay of the sound propagation in the human vocal tract. The speed of sound in the warm and moist air in the vocal tract is approximately 350 m/s, and the average length of the vocal tract for a male is approximately 17.5 cm. The sound is propagating through the tract in $0.175\ \mathrm{m}/350\ \frac{m}{s} = 0.5$ ms. If the model would consist of 8 uniform sections, eight scatterings would take place while the sound wave is propagating through the whole tract, meaning that scattering calculations have to be performed in the tract every $0.5\ ms/8\ = 0.0625$ ms. This corresponds to a sampling frequency of 16 kHz.

Intuitively it would seem reasonable to make the propagation of sound from one end of a section to the other to last one time instant, like in the previous example. The wave would propagate forwards in one section, and meanwhile its reflection backwards in the previous one. However this leads to inefficiency in calculation, and at the end of the tube, in a fully lossless case there would only exist one nonzero output value at every other time instant. [25]

The model is made more efficient when the delay in one segment is half a sample, meaning that two consecutive steps of the full-sample model are combined into one single step. This is possible because, in the full sample model, there is only a forward wave in every other tube at each time instant and a backward wave in every other. When the scattering equations are calculated for the even sections first and after half a delay for the odd sections using the information already known about the previous step, the two steps are effectively combined.

The advantage of the half-sample delay model is that the amount of sections can be doubled when using the same sampling frequency as in the full-sample delay model. If there would be 8 sections, everyone corresponding to a half-sample delay, the sampling frequency would have to be only 8 kHz. The comparison between the full-sample delay model and the half-sample delay model can be observed in Figure 5 and Figure 6.



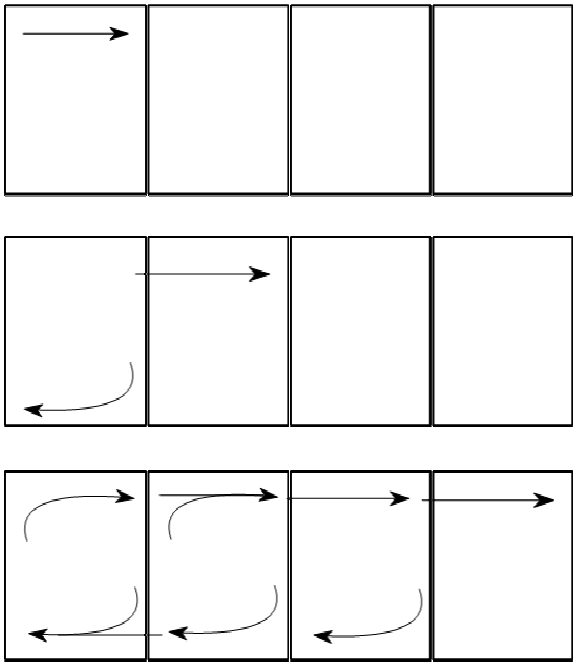
1st sample — Impulse in the first section.

2nd sample — Wave continues to the next section and meanwhile the reflection travels backwards in the previous one.

3rd sample — Forward waves in odd sections and backward waves in even sections.

4th sample — Forward waves in even sections and backward waves in odd sections. First output value is obtained.

Figure 5. Illustrative example of a full-sample delay Kelly-Lochbaum model with 4 sections.



1st sample — Impulse in the first section.

2nd sample — Wave continues to the next section and meanwhile the reflection travels backwards in the previous one.

3rd sample — Now when the even sections are calculated first and after half a delay the odd ones, the result is as steps 3 and 4 of the full sample delay model combined. First output value is obtained.

Figure 6. Illustrative example of a half-sample delay Kelly-Lochbaum model with 4 sections. From third sample onwards the new waves in the tube are combinations of two consecutive samples in the full-delay model.

3.3 Lip rounding in speech production

Length of the lips can be varied in the speech production, as can be observed by comparing vowels /a/ and /o/ for instance. Varying the lip length is usually referred to as *lip rounding*. Lengthening of the lips causes the lowering of all resonance frequencies in the range of the first four formants [23]. Lip rounding can be taken into account in models by lengthening the last section of the vocal tract from its original length. Because of the fixed sampling frequency, accurate measures for the sample values can be obtained only at certain time instants at the tube junctions. If the length of the last section is not a multiple of the distance of the sound wave propagating in a sample interval, a sample value between the existing values has to be approximated.

Delays that are not multiples of the sample interval are called *fractional delays*. Different methods for creating these delays are widely discussed in reference [27]. A fractional delay calls for the use of a sample value lying somewhere between the existing sampling points, which is impossible because of the fixed sampling frequency. Instead, the problem can be solved by reconstructing the continuous bandlimited signal, shifting it in time domain by the desired delay, and then resampling it at the original sampling instants. This is called *bandlimited interpolation*. Figure 7 shows the process of getting the new samples by bandlimited interpolation.

In practical applications, there is no need to go through the whole reconstruction and resampling process, but the operation can be reduced to appropriate linear filtering. There are several different methods of implementing filters that produce fractional delays. In this work, the simplest FIR design called Lagrange interpolator is used. The advantages of Lagrange interpolation are that the calculation of the filter coefficients is easy and fast, it has very good magnitude and phase response at low frequencies, and that the magnitude response never exceeds one, which keeps systems with feedback stable. The coefficients of the filter are calculated using formula

$$h(n) = \prod_{\substack{k=0 \\ k \neq n}}^{N} \frac{D-k}{n-k}, \qquad \text{for } n = 0,1,2,\dots,N \tag{20}$$

where N is the filter order and D is the desired delay.

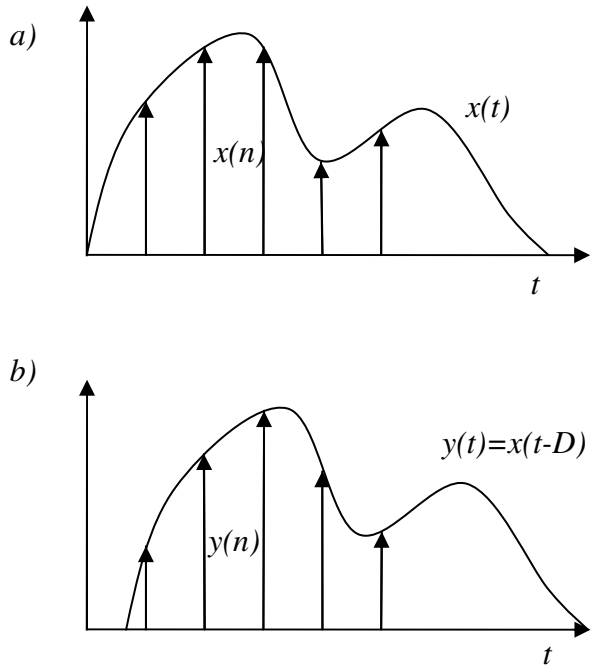


Figure 7. a) Original discrete signal *x(n)* with arrows and reconstructed continuous signal *x(t)* with line. b) Delayed continuous signal *y(t)*, and new sampled signal at original time instants *y(n)*.

In the case of lip rounding, it can be thought that if the lips are lengthened, the signal gets more delay before reaching the lip opening. When the wave propagates back towards the last junction, it has gained a delay equal to twice the time it takes for the sound to travel the length of lip rounding. The fixed sampling frequency causes that a precise value for the reflected volume velocity is needed at the original location of the sample, being the end of the non-lengthened last tube. This can be approximated by calculating the fractional delay using the current sample and a certain amount of previous volume velocity samples at this location.

3.4 Lip radiation impedance

When the sound wave radiates out of the vocal tract at the lip opening, the reflected and radiated wave components are affected by the *lip radiation impedance.* Due to this impedance, the reflection coefficient at the lips is frequency dependent so that higher frequencies are radiated to surrounding environment more strongly than lower ones. Correspondingly, the higher frequencies that are reflected towards the vocal tract are more attenuated. In terms of the volume velocity transfer function of the vocal tract, the lip radiation impedance is a loss that dampens more strongly the higher formants than the lower ones.

This is not exactly the same effect as volume velocity radiation from the lip opening which constructs a pressure wave in the free field around the speaker. This

pressure wave could be estimated by differentiating the radiated volume velocity, which strengthens the higher frequencies approximately 6 dB per octave. The differentiation should be performed when speech is synthesized.

Laine has presented a model for lip radiation impedance in z-domain for digital simulations [28]. In his work, two FIR lip radiation impedance filter models were constructed using radiation impedance approximated by the formula

$$Z_r(\omega) = C \cdot [1 - \cos(\omega T)] + j \cdot B \cdot \sin(\omega T) \tag{21}$$

where the optimum values for the coefficients *C* and *B* are found by minimizing the mean square error, MSE, between the modeled impedance and the actual acoustic impedance using a piston in sphere model [29]. *T* is a factor dependent on the lip area. For digital simulations, the cosine and sine functions are transferred to the z-domain by using the Euler formula

$$\begin{aligned} \sin(\omega T) &= \frac{1}{2j}(z^1 - z^{-1}) \\ \cos(\omega T) &= \frac{1}{2}(z^1 + z^{-1}) \end{aligned} \tag{22}$$

Also, a pole-zero-model for the radiation impedance can be obtained by minimizing the MSE. The transfer function between the radiation impedance and the impedance of the lip section gets a form

$$z_{pz}(z) = \frac{a \cdot (1 - z^{-1})}{1 + b \cdot z^{-1}} \tag{23}$$

At 16 kHz sampling frequency the factors *a* and *b* are found to be

$$\begin{aligned} a &= 0.0779 + 0.2373 \cdot \sqrt{A} \\ b &= 0.8430 - 0.3062 \cdot \sqrt{A} \end{aligned} \tag{24}$$

Where *A* is the area of the lip opening in cm^2.

3.5 Vocal tract shape and resonance frequencies

If the vocal tract is depicted as a uniform tube with a high-impedance glottal source and open end at the mouth, the resonance frequencies of such a tube are represented by standing volume velocity waves as in a *closed-open quarter-wave resonator*. There is always a node located at the glottis and anti-node at the lips. In addition, further nodes and anti-nodes may exist along the vocal tract depending on the frequency of the wave. The first formant frequency corresponds to the standing quarter-wave mode of the tract

and the next standing wave modes correspond to the second, third, etc. formants at frequencies that are 3, 5, and 7 times of the first formant frequency.

The problem of how the constrictions or expansions at different locations along the vocal tract affect the formant frequencies can be approached by examining the process of wave propagation in an ideal uniform tube. For this it is important to understand how the pressure and volume velocity waves act in a uniform tube. The standing waves are caused by the superposition of forward and backward travelling waves of the same frequency and amplitude. At a closed end, the particle velocity at the solid surface has to be zero and the particle velocity at the reflection is reversed. This causes a volume velocity node at the closed end. The phase of the pressure stays the same at the reflection, causing doubling of the pressure compared to the initial wave. This creates a pressure antinode at the closed end. At the open end, the sound pressure collapses and particle velocity increases, causing a pressure node and a velocity antinode. **[**30**]**

During wave propagation, compression of air at a pressure node stores elastic potential energy E_p, which is then transformed into the kinetic energy, E_k, of the unit volume when the air expands. Potential energy is proportional to the squared pressure, and kinetic energy is proportional to the squared volume velocity [30]. The volume velocities, pressures, and potential and kinetic energies for the first 4 resonance modes are shown in Figure 8.

The sensitivity function of a formant is defined as the difference between the kinetic and potential energies E_k - E_p. At the maxima of the sensitivity functions potential energy is zero and kinetic energy is at its maximum. Decreasing the cross-sectional area of the tube at these points decreases the volume velocity, leaving the pressure unchanged. If the acoustic resonance is considered to be produced with a cascade LC-resonator, decrease in the volume velocity (current) means increase in the inductance. According to the general LC-resonator formula

$$f = \frac{1}{2\pi\sqrt{LC}} \tag{25}$$

the resonance frequency is lowered. At the locations of all positive values of the sensitivity function, constriction causes increase in inductance and decrease in capacitance in proportion to the values of volume velocity and pressure at the given location[3]. When the sensitivity function is positive, the increase in inductance outweighs the decrease of capacitance, causing the resonance frequency to lower. Constrictions directly at zero-crossings keep the resonance unchanged due to annulling change in both reactances. Correspondingly, at negative values of the sensitivity function the resonance frequency is raised. The sensitivity functions for a uniform tube are shown in Figure 9. [31]

[3] Note that in a lossless tube the inductance per unit length is inversely proportional to the cross-sectional area meanwhile the capacitance is directly related to the area. This gives some intuition to interpret (25). When, e.g., the constriction occurs at a place with low volume velocity the increase in inductance has no effect meanwhile the pressure at the point may be high enough to cause an effect due to the decreasing capacitance.

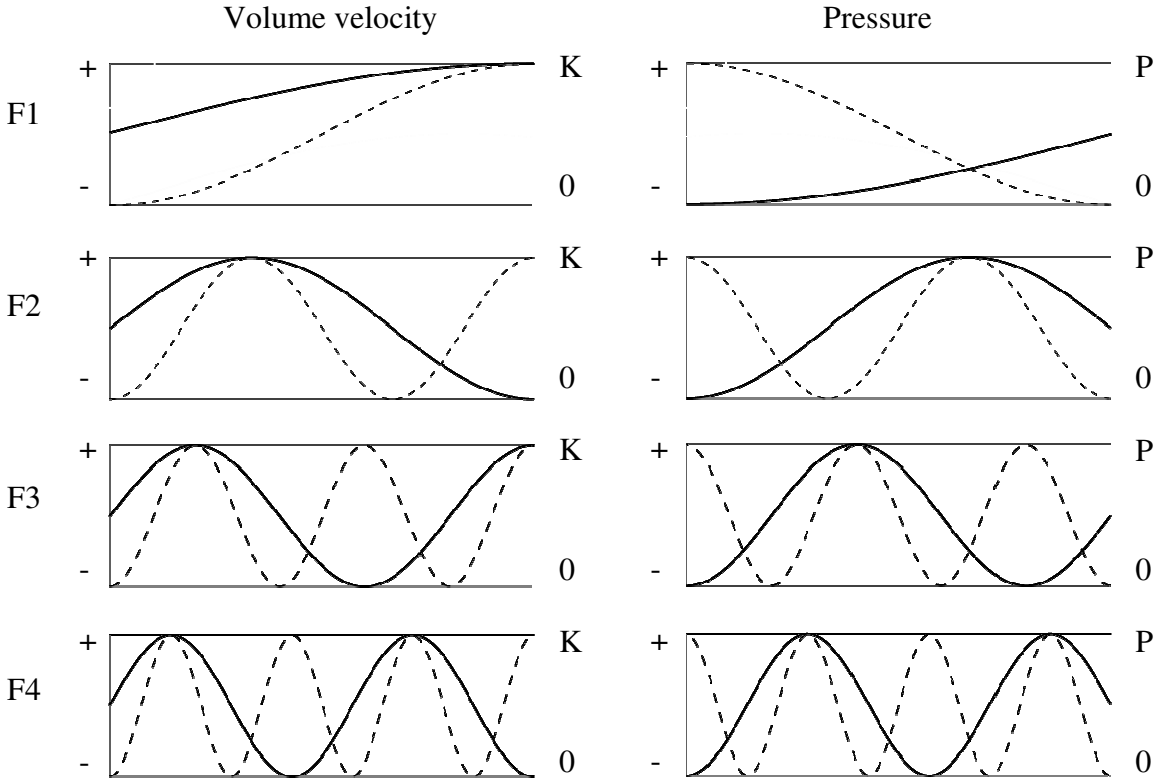


Figure 8. Continuous line shows standing wave patterns of the first four formant frequencies in an ideal uniform tube for volume velocity (left) and pressure (right). Dashed line indicates the corresponding kinetic energy on left and potential enegy on right. The minimum value for these energies is 0. Closed end of the tube is at the glottis and open end at the lips.

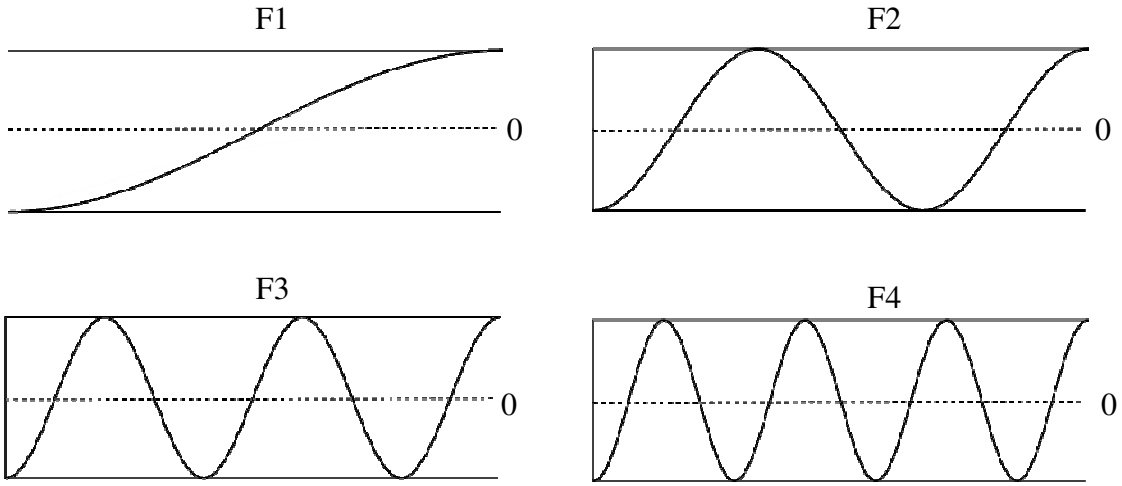


Figure 9. Sensitivity functions for the first four formants in the uniform vocal tract.

3.5.1 Formant sensitivity functions in terms of cosine series

Fourier series is a mathematical method of decomposing a continuous periodic function into a sum of sine and cosine functions. Fourier series of a 2π-periodic function takes the form

$$f(x) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty}[a_n \cos(nx) + b_n \sin(nx)] \qquad (26)$$

where

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(nx)\,dx\,, \qquad n \geq 0 \qquad (27)$$

$$b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(nx)\,dx\,, \qquad n \geq 1 \qquad (28)$$

The functions can be interpreted so that the function *f*(*x*) can be implemented in terms of an infinite amount of cosine and sine terms that take the weighting factors as shown in the two latter equations. If all the sine terms are zero, the function takes the form that is called *cosine series*. From Figure 9 it is observed that the sensitivity functions are indeed terms of a cosine series. The sensitivity functions for the first four formant frequencies are the cosine series terms for the wavelengths $\frac{1}{2}\lambda$, $\frac{1}{6}\lambda$, $\frac{1}{10}\lambda$ and $\frac{1}{14}\lambda$, corresponding to the odd values of *n* in the equations. λ denotes the wavelength of the first resonance.

It is clear directly from the definition of Fourier and cosine series that the terms are *orthogonal*, meaning that a change in the weighting factor of one term does not affect the other terms. This means that perturbations in the tract according to each of the four sensitivity functions will only move the formant frequencies corresponding to them. Furthermore, if the entire vocal tract shape is represented in terms of cosine series, the creation of changes only in these four terms should move the first four formant frequencies in a desired manner.

Naturally, this kind of a method is an approximation since the standing wave pattern in the vocal tract changes as soon as a change in the vocal tract shape occurs. Also, this method works as mentioned for the uniform tube only, because the standing wave pattern is simple and known accurately only for that shape. When the vocal tract has more complex shapes, the waveforms in the tract are different, and the calculation of the sensitivity functions becomes more complicated. [23]

However this method gives relatively good results also for more complex vocal tracts. Although the change in the formant frequencies is not truly orthogonal anymore, the behaviour of the resonances can still be fairly well estimated with proper methodology. This was noticed while experimenting with the vocal tract changes during this work. More detailed description of the experiments can be found in section 5.3.

Sensitivity functions for varying tract shapes can be calculated also as introduced by Fant and Pauli in [32]. After every change to a vocal tract shape, a new sensitivity function could be calculated according to which the new change is performed. During the vocal tract shape variation, this would lead to extra calculation between every iteration. Due to the vast amount of vocal tract variations performed in this work, an adequate series of vocal tract shapes can be obtained with the less computational effort using the variation method based on the sensitivity functions of uniform tubes.

3.6 Linear prediction for formant analysis

Linear predictive analysis is widely used in speech research due to its ability to produce accurate estimates of the spectral characteristics of speech and its small computational cost. The spectral qualities of sounds can be represented efficiently by linear prediction, using only a low number of parameters. In this work, linear predictive analysis is used to extract formant frequencies from the impulse responses obtained by the vocal tract model, as well as the recorded speech signals.

The idea of linear prediction lies in the source-filter model of the speech production, and is described well in the reference [33]. The vocal tract in a short time analysis is considered as a linear, time invariant, all-pole system that creates resonance peaks at certain frequencies. In general, such a system takes the form

$$H(z) = \frac{S(z)}{E(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{29}$$

where G is the gain parameter and $\{a_k\}$ are the *system filter coefficients*, known also as *vocal tract filter*, or *synthesis filter coefficients*. The filter order is p. Linear predictive analysis estimates these coefficients from a speech signal.

In discrete time domain, the speech samples $s[n]$ can be calculated from the p previous samples using the formula

$$s[n] = \sum_{k=1}^{p} a_k s[n-k] + Ge[n] \tag{30}$$

where $e[n]$ is the excitation signal. The predicted speech using the linear predictor takes the form

$$\tilde{s}[n] = \sum_{k=1}^{p} \alpha_k s[n-k] \tag{31}$$

meaning that the new sample can be predicted from p previous samples multiplied with the *prediction filter coefficients* $\{\alpha_k\}$. The error between the real sample and the estimated sample is called *prediction error* and is defined as

$$d[n] = s[n] - \tilde{s}[n] = s[n] - \sum_{k=1}^{p} \alpha_k s[n-k] \tag{32}$$

The prediction error is thus an output of a FIR filter with a z-domain representation

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} \tag{33}$$

If the predictor coefficients $\{\alpha_k\}$ would match the system coefficients $\{a_k\}$ perfectly, the prediction error filter $A(z)$ would be an ideal filter for the system, yielding

$$H(z) = \frac{G}{A(z)} \tag{34}$$

The main challenge in this method is now minimization of the average prediction error in a certain time window by optimizing the prediction filter coefficients. There are two commonly used methods for solving this, the *covariance method* and the *autocorrelation method*. The autocorrelation method is more widely used because of the guaranteed stability of the resulting filter and its easy solution by using *Levinson-Durbin-recursion*.

The minimization of the prediction error takes the form

$$E = \sum_{n=-\infty}^{\infty} e_n^2 = \sum_{n=-\infty}^{\infty} \left(s[n] - \sum_{k=1}^{p} \alpha_k s[n-k] \right)^2 \tag{35}$$

which is a *least squares* problem, since there are more equations than variables to be solved. Here it is assumed that an infinite amount of signal values, $s[n]$, are available. The equation is solved by setting the partial derivatives to zero with respect to the prediction coefficients

$$\frac{\partial E}{\partial a_i} = 0, \;\; 1 \leq i \leq p \tag{36}$$

leading to equation

$$\sum_{k=1}^{p} \alpha_k r(i-k) = r(i), \qquad 1 \leq i \leq p \tag{37}$$

where $r(i)$ are the autocorrelation terms

$$r(i) = \sum_{n=-\infty}^{\infty} s[n]s[n-i] \ , 0 \leq i \leq p \tag{38}$$

In practical applications, the signal is not infinite but *windowed*, meaning that only a short period of the signal is chosen for examination at a time. The formant frequencies concerning individual sounds will be obtained when the signal is examined during a time period inside which approximately one pitch period is captured. The window is usually chosen to be somewhat longer than the expected pitch period to gain reliable formant information even if the window does not exactly meet the location of the main excitation at the glottal closure. Typical window length to capture a few periods of voiced speech is around 20-30 ms. Because the autocorrelation function is even and due to windowing with a window of *N* samples, the previous equation inside the window takes the form

$$r(i) = \sum_{n=i}^{N-1} s[n]s[n-i] \ , 0 \leq i \leq p \tag{39}$$

When all the required autocorrelations are calculated, equation (37) can be written in a matrix form

$$\mathbf{R}\boldsymbol{a} = \boldsymbol{r} \tag{40}$$

The values of the matrix **R** are constant across any diagonal**,** and it is thus a so-called *Toepliz matrix*. This property leads to an easy inversion with the already mentioned Levinson-Durbin-recursion, and the prediction coefficients are solved as $\boldsymbol{a} = \mathbf{R}^{-1}\boldsymbol{r}$. The gain factor *G* only shifts the resulting spectrum to match the original spectral energy and is calculated as

$$G = \sqrt{r(0) - \sum_{k=1}^{p} a_k r(k)} \tag{41}$$

The frequency-domain interpretation of linear prediction is that the LP tries to model the resonance peaks of the corresponding short-time Fourier transform. The amount of smoothing compared to STFT is controlled by the choice of the model order. Generally speaking, low model orders find only the highest resonance peaks whereas higher order models are also able to model less energetic resonance at higher frequencies. From the acoustic theory of speech production it is known that for a male speaker, the spectrum of the vocal tract filter has about one resonance peak per kilohertz and the glottal pulse and radiation can be modeled with two additional complex pole pairs. In general, a good rule of thumb in speech analysis is to choose an LP model of order of $p = 4 + f_s/1000$ to get all the essential information of the speech signal. [33]

In Figure 10, the spectrum of an extracted segment of a Finnish speech signal during a vowel sound /a/ is shown with 2048 point FFT and an order 20 linear prediction model. 16 kHz sampling frequency was used in the recording. Frequencies are shown on a scale from 0 to 4 kHz, which is the band in which the first four formant frequencies are expected to lie. It can be seen that the highest frequency modes are accurately modeled and the unwanted small variation is not included in the representation.

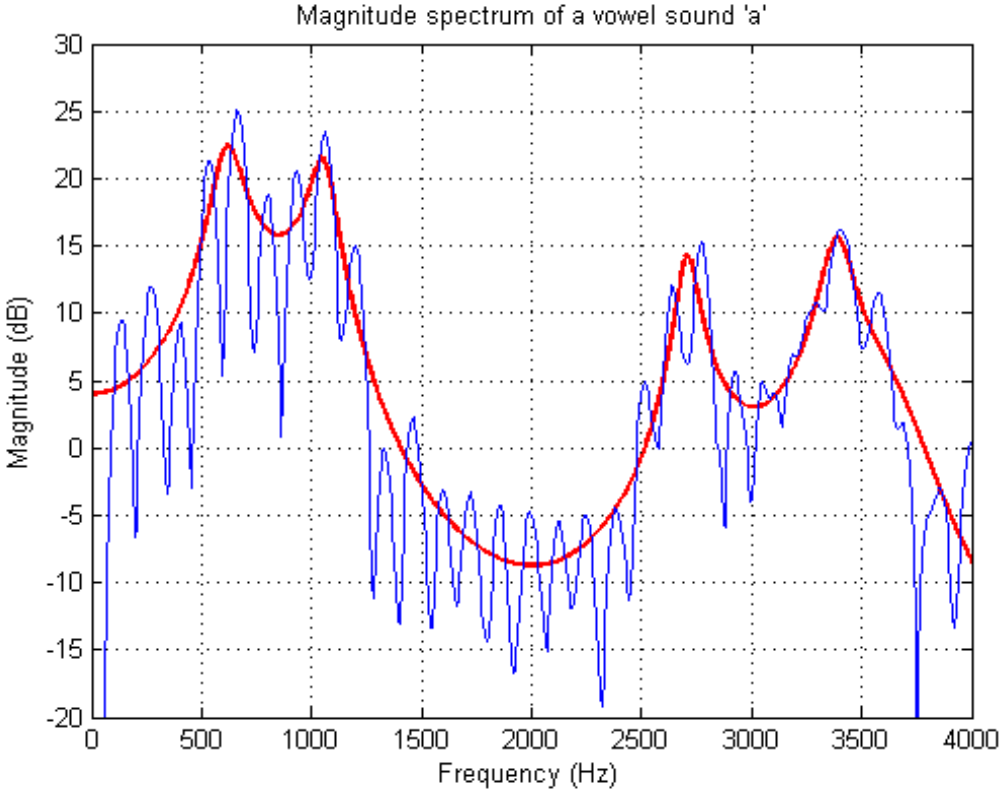


Figure 10. Frequency spectrum of a Finnish vowel /a/ with FFT (thin line) and an order 20 linear prediction model (thick line). The formant frequencies are easily extracted from the highest peaks of the LPC-curve.

3.6.1 Estimating formant frequencies using linear predictive analysis

Since the poles of the system filter are the roots of the prediction filter polynomial $A(z)$, the resonance frequencies can be solved from the complex roots of the polynomial. The poles of the system filter are located inside the unit circle with magnitudes varying between zero and one. The most significant amplitude peaks corresponding to the formant frequencies have also the highest pole magnitudes. When a high prediction order is used, some poles are also located to frequencies in between the actual formant frequencies of interest. These extra poles generally have a smaller magnitude compared to the poles corresponding to the formant frequencies.

To avoid the situation where the extra poles could be chosen as formant frequencies, a threshold magnitude for finding the actual formant frequencies from the system filter poles is set. After experimenting with recorded speech, a threshold factor

of 0,9 was chosen in this work. Above this threshold the formant frequencies could be found with all recorded vowel sounds as well as with the vocal tract impulse responses. When the poles with magnitudes more than 0.9 are extracted out from all the poles, and then sorted in ascending order according to the phase angles, the four first formant frequencies can be easily selected.

4 Implementation of a lossy Kelly-Lochbaum model

This section describes how the vocal tract model is constructed. Reasoning for the choice of variables is discussed as well as description on how the theoretical background explained in chapter 3 is used in practice. The experiments reported in this work were performed using the Matlab® environment. At this point, in addition to the theoretical part, the actual implementation and simulation stage of this master's thesis begins.

The lossy Kelly-Lochbaum vocal tract model is implemented for the purpose of acoustic-to-articulatory mapping. The vocal tract model can be used to obtain acoustic data concerning a certain vocal tract area function. As parameters, the vocal tract model takes the cross sectional areas of the 16 uniform sections and the length of lip rounding, and produces an impulse response of the vocal tract as output. The model is used later to synthesize all the sounds corresponding to the vocal tract area functions created in the variation phase.

4.1 Calculation of the reflections at tube junctions

Kelly-Lochbaum vocal tract model was implemented using the principles described in sections 3.2.1 and 3.2.2. The volume velocity model was used, since the glottal excitation is often modeled as volume velocity wave. Using excitation based on volume velocity variations originates to the assumption of constant subglottal air pressure, causing the corresponding volume velocity to be proportional to the area of the glottis. The varying area at the glottis again is easy to model. [21]

The frequency response of the vocal tract model was obtained in this work by computing the volume velocity impulse response of the KL model. This is performed by introducing a volume velocity impulse (only one sample) excitation at the glottal end. Then the wave propagation with reflections is calculated throughout the whole non-uniform tract. If the glottis is chosen to situate on the left and lips on the right, the forward volume velocity wave propagates from left to right. At the same time instant, the scattering equations are used in each junction to calculate the new values for the forward and backward waves in all the sections. The way the model was implemented in this work is the half-sample delay model as introduced in chapter 3.2.3, but it was calculated in a slightly more efficient manner. The calculation is performed from the lips towards glottis, because this way the backward travelling wave values do not need

to be stored for each section individually. When the calculation goes backwards section by section, the new junction always uses the backward volume velocity value obtained in the previous scattering. Because of the backward computation, the reflection coefficients are also conveniently calculated from lips towards glottis. The signs of the reflection coefficients in the volume velocity model (Figure 4) are thus changed.

When all of the junctions are processed, a constant reflection coefficient is applied at the glottis to calculate the new forward travelling wave in the first section. Before the beginning of scattering calculations at a new time instant, a backward reflecting volume velocity value is calculated from the lips, as explained later in chapter 4.3.

4.2 Selection of model parameters

Speech synthesis can be made using an endless variety of models with varying amount of parameters. For acoustic-to-articulatory mapping purposes, a model that reaches a wide repertoire of different speech sounds, but works with as few parameters as possible, is preferred for computational efficiency and memory requirements. In this work, the goal was to fill the formant space with mapped points as widely as possible in order to support the estimation of the vocal tract shape trajectories when entering vowel sounds from unvoiced sounds, as well as during transitions between two different vowel sounds. Also, in order to limit the parametric variation to realistic vocal tract shapes, the variation was especially created around the known area functions of the eight Finnish vowels. This was done by adjusting the vocal tract shape using terms of cosine series, creating smooth variations of the original vowels and avoiding physiologically unfeasible vocal tract shapes.

4.2.1 Vocal tract shape approximation by a limited number of tube sections

For the Kelly-Lochbaum vocal tract model, a fixed number of uniform sections needs to be decided. The model should be able to give good results in synthesizing vowel sounds but it should also be compact enough to keep computation time reasonable. Also, choices on how the cross-sectional areas of the uniform sections are selected from a corresponding physiological continuous tract have to be made.

There are some measurements made by MRI and X-ray imaging that show accurate physiological data regarding the area functions in tens of points throughout the vocal tract during production of a vowel. It would be satisfying to be able to convert these continuous area functions into a number of uniform tubes of a fixed length and simultaneously maintain the characteristics of the vowel sound produced. There are different choices on how this conversion could be made. In Flanagan's early model, the mean of the model areas in the limits of the section was chosen as the area of the section [7]. Another way is to keep the volume inside the original shape and the discrete section constant.

Local constrictions are often the most important factors in how the vowels sound. This can be already seen from the basic classification of vowel sounds, since the vowels are often classified due to their degree of constriction and tongue hump position (front,

central, back) [21]. If any kind of averaging in the model is used, the degree of highest constriction could easily be averaged out, causing a less realistic vowel sound. This could lead to the need of manual adjustment of some tube areas to match the highest constriction and get a good vowel correspondence.

Since the sensitivity functions are terms of cosine series, and the vocal tract shapes are to be varied using the cosine coefficients, it was chosen to describe the area function in terms of cosine series also. For example, if the measured area function has area information at N points along the tract, there would be a requirement for N uniform sections to preserve all the necessary information, and this would need also N terms of cosine series to obtain perfect reconstruction. A good way to get a smooth correspondence for a smaller amount of tubes, P, is to "lowpass filter" the original vocal area function, conserving only the frequency components that can be represented by the amount of sections chosen. This is done by calculating a discrete cosine transform, DCT, of the original shape, resulting into N coefficients from which only the first P coefficients are stored.

For visual convenience, a smooth profile of more points can be created from the DCT coefficients by zero padding the DCT vector to a desired length and calculating the inverse cosine transform. The process explained is illustrated in Figure 11. The original shape is obtained from the reference [34] and corresponds to an English vowel /Λ/ as in word "ton".

This method is similar to averaging, so the problem of averaging out some important components is not completely solved. However, in this study the manual adjustment of the tract shape is required in the beginning in any case. This is because exact area functions for the Finnish vowel sounds do not yet exist in the current literature.

One might argue that rather the profile smoothed with zero padding in the undermost picture of Figure 11 could be discretized to get the final tube areas. It can be seen that for example the last lip section would be approximated somewhat better from the smooth profile. This is due to the fact that zero padding does not add any extra information, but increases the number of points by interpolation. Discretizing this smooth shape differently might model some components better but some also worse. In the case of the figure presented, the area of the lip section would correspond better to the original area at the lips, but the area of the highest expansion would be modeled lower than the original. In order to keep the model compact and the cosine transformation matrices small, the first P cosine series terms from all area functions from the literature are used in this work.

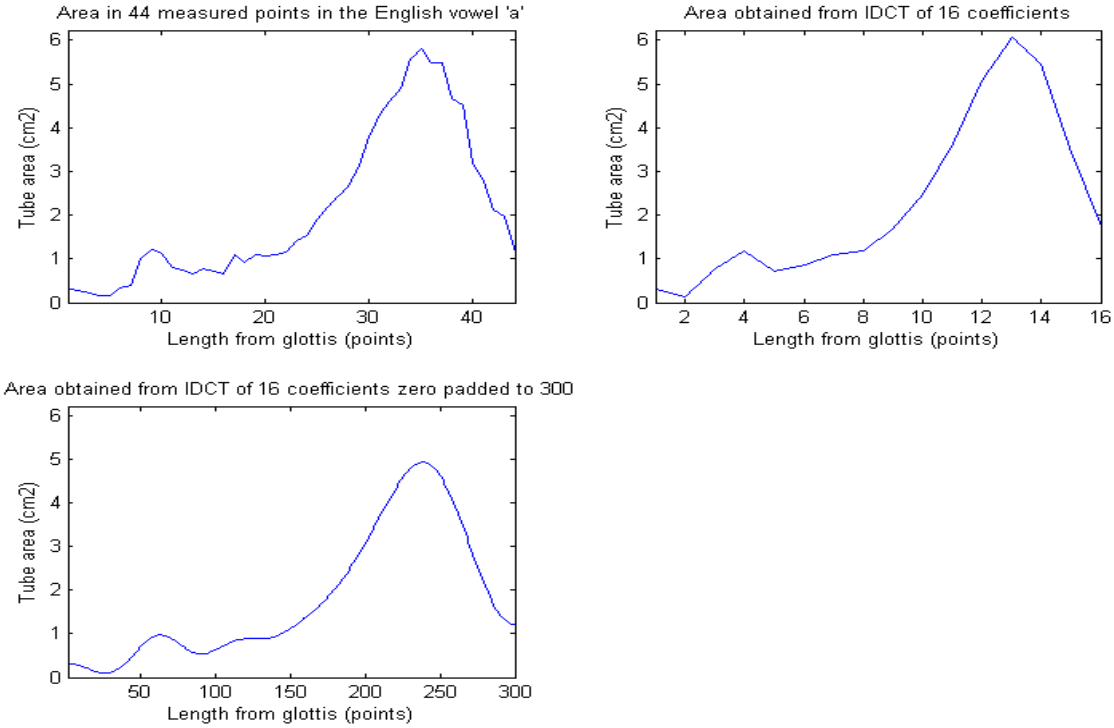


Figure 11. Discretization of a measured vocal tract shape to 16 uniform sections using the discrete cosine transform. Measurement of a vocal tract area in 44 points (up left). 16 areas obtained by taking the DCT of the original shape and then IDCT for the first 16 conserved coefficients (up right). Smooth area function from zero padding the DCT to 300 samples and taking IDCT (down).

4.2.2 Selection of the number of sections

When experimenting with vocal tract models during this work, the Kelly-Lochbaum model was created using only 8 uniform sections at first, corresponding to a sampling frequency of 8000 Hz. When the areas of the tubes were chosen properly and a good glottal excitation pulse was created, the vowel sounds sounded realistic and the spectra calculated from the impulse responses of the model were as expected for vowel sounds. Nevertheless, this choice for the number of sections had some drawbacks. First of all, when the shape of the real physiological vocal tract has to be represented with only eight uniform tubes, each being approximately 2.2 cm in length, the physical correspondence of the model suffers. Realistic vowel sounds can be still created by finding optimal areas manually by trial and error and listening to the results at the same time. Although this gives a decent overall view on the vocal tract shape, it does not provide very accurate information concerning some steeper local constrictions being less is distance than the tube section length.

Another drawback was the sampling frequency of 8000 Hz. When the lip radiation is modeled, the original work by Laine [28] states that the pole-zero model described here at section 3.4 works down to sampling frequencies of 14 kHz. In order to avoid the construction of a more complicated filter, it is possible to resample the output with a sampling frequency of 8 kHz to 16 kHz to be used by the pole-zero model. Resampling

could be done using similar Lagrange interpolation as in lip rounding. This would cause the output to have twice the original number of samples, from which there is a need to perform resampling again in order to obtain the backward wave reflected from the lips back into the vocal tract model working at the 8 kHz sampling frequency. If the lip rounding is modeled simultaneously, complicated interpolation structures have to be created. During the development it quickly became clear that in order to keep the model as simple as possible, additional 8 tube sections should be added, increasing the sampling frequency to 16 kHz. This doubles the computation time needed by the scattering calculations, but computation is also reduced by simplifying the lip radiation model to work without resampling.

The use of 16 sections also gives a better physiological correspondence, reduces the problem of modeling the highest constrictions, and later gives a more accurate shortest trajectory finding algorithm through the possible vocal tract shapes when examining continuous changes in speech signals.

4.3 Implementation of lip radiation impedance and lip rounding

The lip radiation impedance is implemented using the pole-zero type model as described in section 3.4. It has to be noted that the filter models the lip radiation impedance the volume velocity wave meets when leaving the vocal tract. The model leads to frequency dependent reflection at the lips. Higher frequencies are better radiated and less reflected. Therefore the higher formants are more strongly damped than the lower ones. The reflection coefficient is obtained from the impedances as shown in equation (18). Now taking into account the reversed direction of calculation, the reflection coefficient for the lips becomes

$$R_{\mathrm{rad}} = \frac{Z_{\mathrm{end}} - Z_{\mathrm{rad}}}{Z_{\mathrm{end}} + Z_{\mathrm{rad}}} \tag{42}$$

where Z_{end} is the impedance of the last section and Z_{rad} the lip radiation impedance

$$Z_{\mathrm{rad}} = Z_{\mathrm{end}} \frac{a \cdot (1 - z^{-1})}{1 + b \cdot z^{-1}} \tag{43}$$

obtained using the transfer function in equation (23). Combining these equations, the reflection coefficient becomes

$$R_{\mathrm{rad}} = \frac{(1 - a) + (b + a)z^{-1}}{(1 + a) + (b - a)z^{-1}} \tag{44}$$

Now the output volume velocity of the tract after the radiation reflection coefficient becomes

$$U_{\text{out}} = (1 + R_{\text{rad}}) \cdot U_{\text{end}} = \frac{2U_{\text{end}} + 2b \cdot U_{\text{end}} \cdot z^{-1} - (b - a) \cdot U_{\text{out}} \cdot z^{-1}}{1 + a} \quad (45)$$

and the backward travelling volume velocity for the last tube

$$U_{\text{back}} = -R_{\text{rad}} \cdot U_{\text{end}} = \frac{+(b + a) \cdot U_{\text{end}} \cdot z^{-1} - (a - 1) \cdot U_{\text{end}} - (b - a) \cdot U_{\text{back}} \cdot z^{-1}}{1 + a} \quad (46)$$

A flow chart of the frequency dependent reflection filter at the lips can be seen in Figure 12.

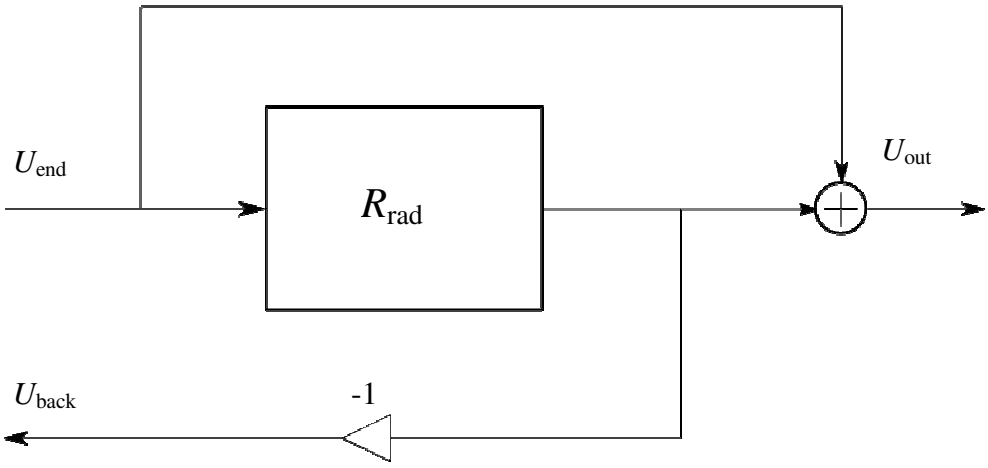


Figure 12. Frequency dependent reflection at the lips realized by a linear filter.

The effect of lip rounding is implemented as a fractional delay as described in section 3.3. In the given implementation, there is a vector for storing five consecutive output values from the tract. The vector is rotated right at every time instant and the new value is added to the first element. Now it can be thought that if the lip rounding length is zero, the required extra delay is also zero, and the reflection for the backward wave is calculated exactly from the first element of the vector that corresponds to the consecutive output signals. If the length of lip rounding is half of the length of one uniform section, the delay for the backward wave is one full sample, and the backward wave is calculated from the second position, corresponding to the previous output value. Any delay in between can be calculated by fitting a Lagrange polynomial to all the five previous samples and picking the value at a desired location related to the lip rounding. In practice, the result is obtained by calculating the fourth order Lagrange interpolation filter coefficients with desired delay using equation (20) and performing a vector multiplication between the time dependent output values and the filter coefficients.

Figure 13 shows the effect of lip radiation impedance. When compared to the spectrum with the constant loss factor at the lips (solid line), the higher formant frequencies are decreased. This confirms the theoretical principles described in section

3.4. This effect corresponds the inductive end correction known in acoustic theory of pipes. Thus the inductive component of lip radiation impedance approximates the end correction.

The effect of lip rounding is shown in Figure 14. The original extra length due to lip rounding was 0.1 times the section length. The dashed line shows the same vocal tract shape with lip rounding of 0.7 times the section length. This causes the formant frequencies to drop and the vowel starts to remind more the Finnish vowel /o/.

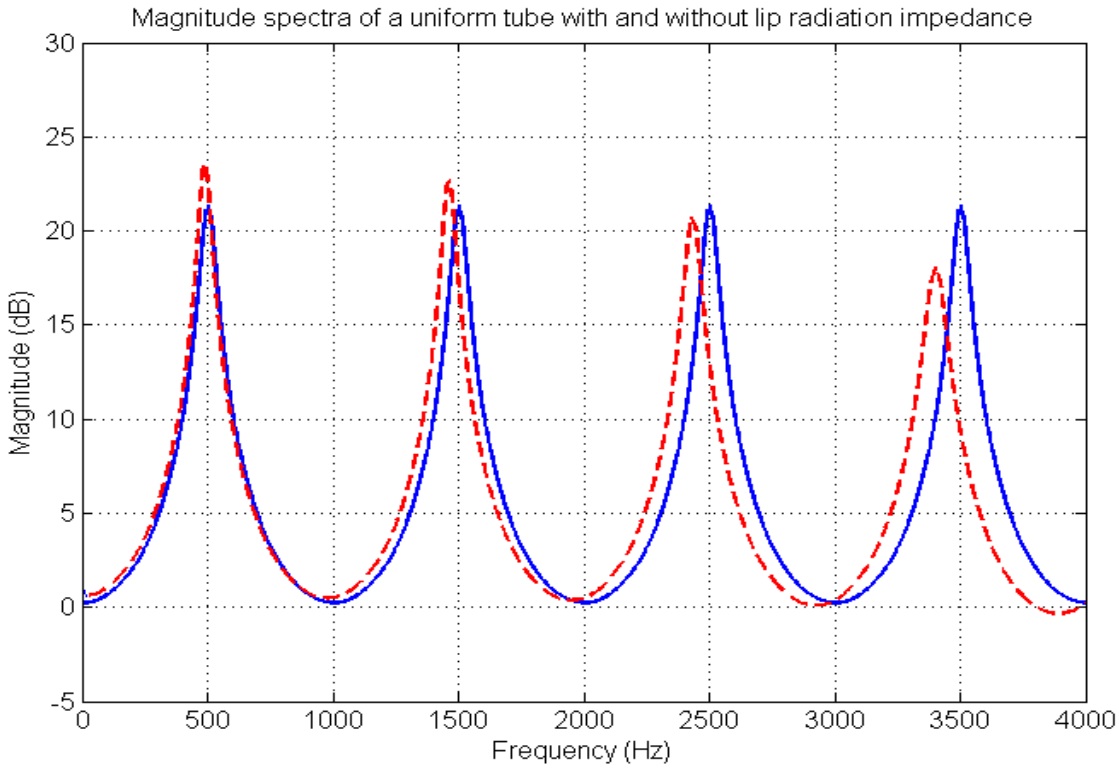


Figure 13. The effect of the lip radiation impedance illustrated with a uniform tube. Solid line shows the spectrum with a constant lip loss factor. Dashed line shows the spectrum with the frequency dependent reflection coefficient.

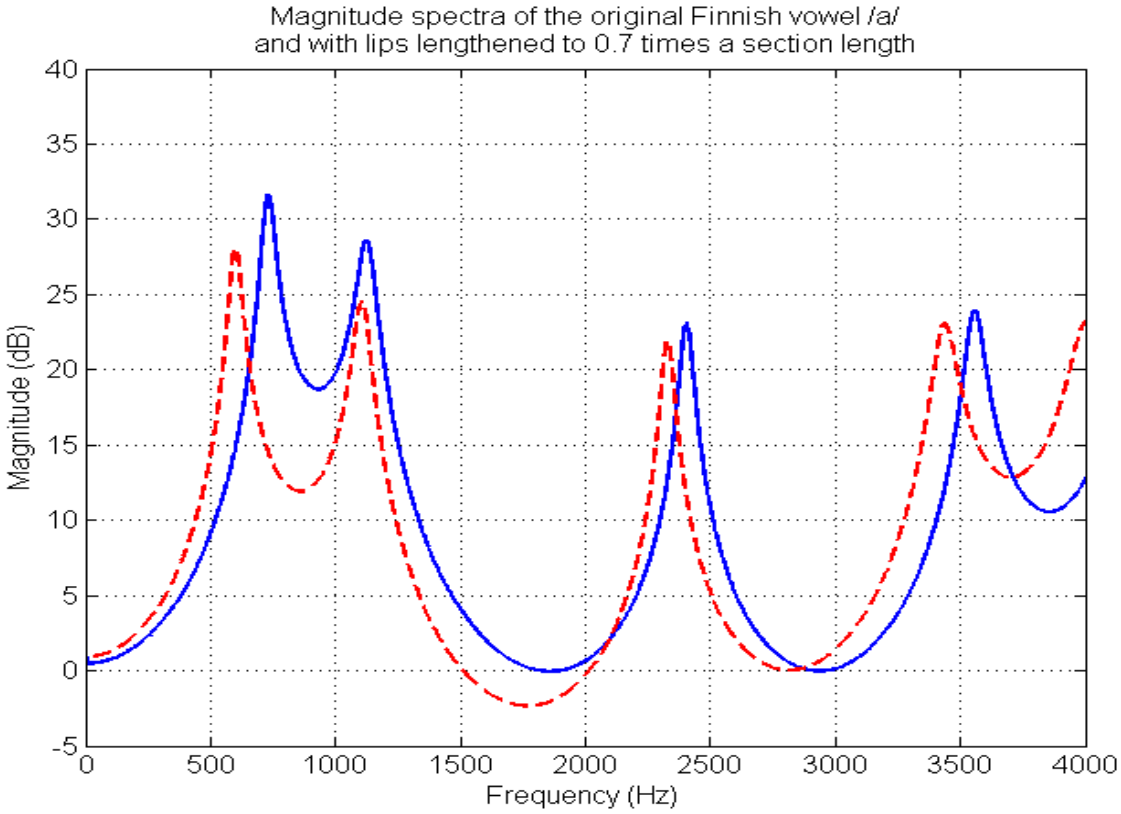


Figure 14. The effect of lip rounding. The solid line shows the original situation of vowel sound /a/ when lips are lengthened by 0.1 times the section length. Dashed line shows the lips lengthened by 0.7 times the section length.

5 Creating the acoustic-to-articulatory lookup table

In this section, a description of a method for the acoustic-to-articulatory mapping is given. Finding of the so-called anchor tract shapes for Finnish vowels is discussed as well as the development of the variation methods used to create reasonable changes to the anchor shapes. Different methods for varying the vocal tract shapes are experimented and the results described. Finally, the created lookup table is visualized and analyzed by plotting a formant chart and a density map.

5.1 Basic idea of acoustic-to-articulatory mapping

The idea of the acoustic-to-articulatory mapping is to find candidates for vocal tract shapes from recorded speech signals. In this study, the mapping is performed by using a lookup table, created by simulating speech production with the compact Kelly-Lochbaum vocal tract model introduced in the previous section. Since modeling of the vowel sounds is emphasized in this work, the creation of the lookup table is started from eight selected vocal tract shapes that correspond to the eight primary Finnish vowel sounds. These shapes are defined as *anchor points*, and smooth changes are introduced to them according to the modulation method to be described.

A variety of vocal tract shapes is created and corresponding formant information stored. The aim is to find vocal tract shapes corresponding to every physiologically possible combination of formant frequencies. This is referred to as *filling in the formant space*. When the amount of variations to the anchor shapes is increased, formation of *clusters* of points around the points corresponding to the anchor shapes can be observed in the F1-F2-plot. When the variation is wide enough, the clusters begin to overlap and finally the whole formant space is filled.

The information stored at each variation is the cross-sectional areas for each 16 uniform sections, the length of lip rounding, and the four first formant frequencies. A test for the physical correspondence of the results is later made in chapter 6 by analyzing spoken test signals. The signal is windowed and a few closest formant equivalences are found for each window. These are converted into vocal tract shapes using the obtained lookup table, and the physical correspondence concerning the smooth movements of the articulatory system is guaranteed by finding a minimum Euclidean distance path through all the choices for possible articulatory shapes. This path is called *shape trajectory*.

5.2 Anchor vocal tract shapes for Finnish vowel sounds

The anchor points for the vowels are selected by manually finding a match between the formant frequencies of the Finnish vowel sounds by Wiik [35] and the corresponding vocal tract shapes obtained from real physiological articulatory measurements. Since exact measurements from Finnish vocal tract area functions do not yet exist, this task proved to be somewhat challenging. Fant has listed area functions obtained via X-ray

imaging for six Russian vowels [23]. For Swedish vowels, MRI imaging has been used to find area functions in the work of Ericsdotter [36]. For an English male speaker, 12 vowel sounds, 3 nasals, and 3 plosives have been examined and accurate vocal tract area functions have been obtained by MRI imaging [34]. None of these studies has been able to record the speech signal produced simultaneously with the vocal tract shape measurement due to the strong impulsive noise generated, and problems with microphone techniques in the intensive magnetic field related to the imaging techniques.

Since the vocal tract shapes for the English vowels seemed the most accurate and had more than 40 measurement points through the vocal tract, these were selected to form the starting point. These shapes were then manually modified to match the Finnish vowel sounds. The article by Story, Titze and E. A. Hoffman [34] provides the spoken words from which the vowel part is examined, so the intuitively closest matches for the Finnish sounds were selected. As described in section 4.2.1, all the area functions were discretized into 16 uniform sections using discrete cosine transformation of the original shape, selecting the 16 first coefficients and performing an inverse transformation back to an area function. Then the eight vowel sounds giving quite a good match to the corresponding English vowel sounds were manually adjusted to match the Finnish vowels. Attention was paid to the minimum constriction of the tract in the original data. If the section seemed to be radically averaged to become less constricted, the area of the section was decreased manually. Using a simple glottal excitation pulse, the adjustment process was carefully listened to and good correspondences for the Finnish vowels were obtained iteratively. The lip rounding was approximated by examining the lip length in corresponding Finnish vowel sounds. The final area functions for the anchor shapes are shown in Figure 15. Lip rounding can be seen as the area exceeding the 16^{th} section. The spectra for all the vowels are shown in Figure 16.

In Figure 17, the first and the second formant frequencies of the resulting vowel sounds are plotted on top of a F1-F2-plot for Finnish vowels as introduced by Wiik [35]. It is seen that the first and the second formants are situated inside the circles drawn in the figure, giving the starting shapes for the variations a good spectral correspondence. This formant plot was also used as an aid for the adjustment of the vocal tract shapes.

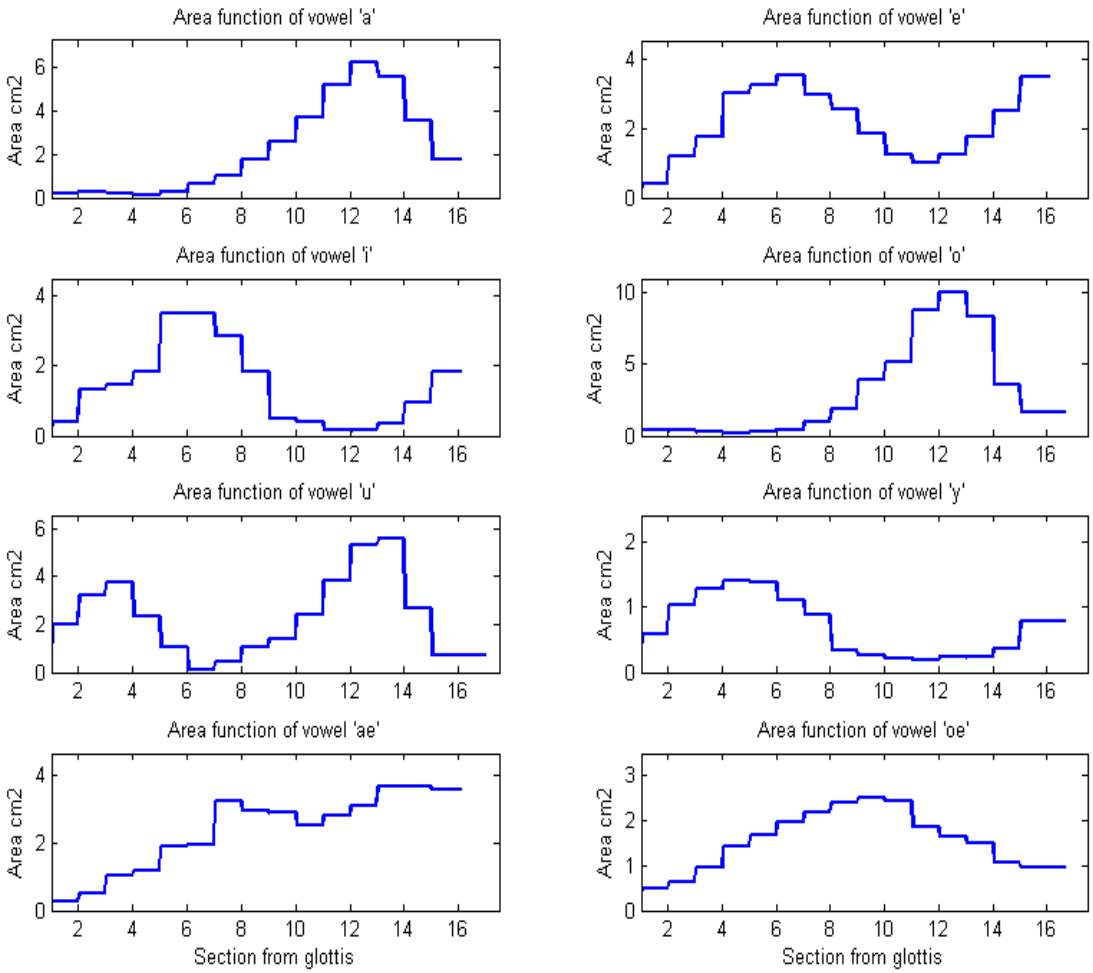


Figure 15. Area functions for the Finnish vowel sounds used as anchor shapes. The area exceeding the 16th section represents the lip rounding.

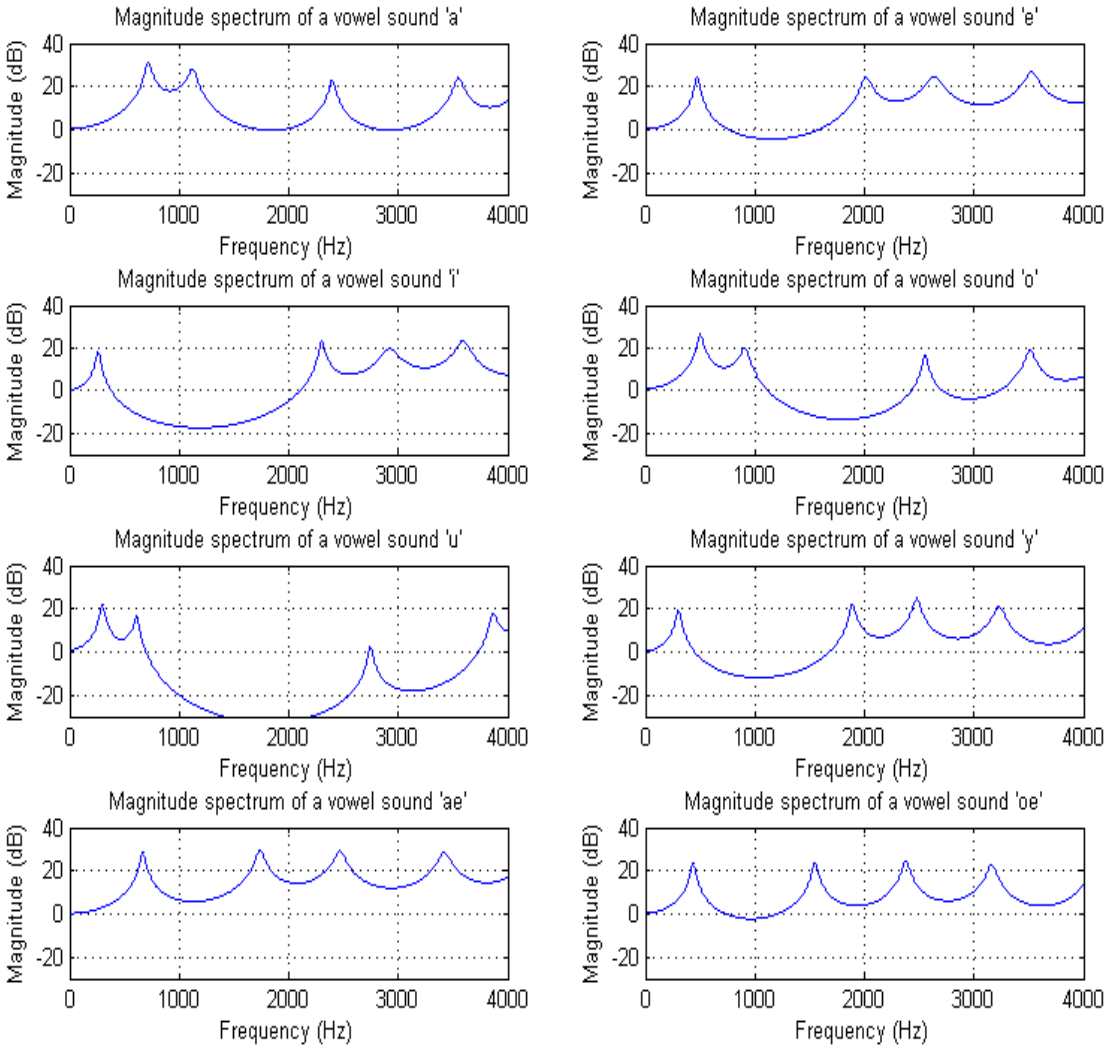


Figure 16. Spectra for all Finnish vowel sounds obtained by adjusting the closest English vowels manually and simulated with the KL model.

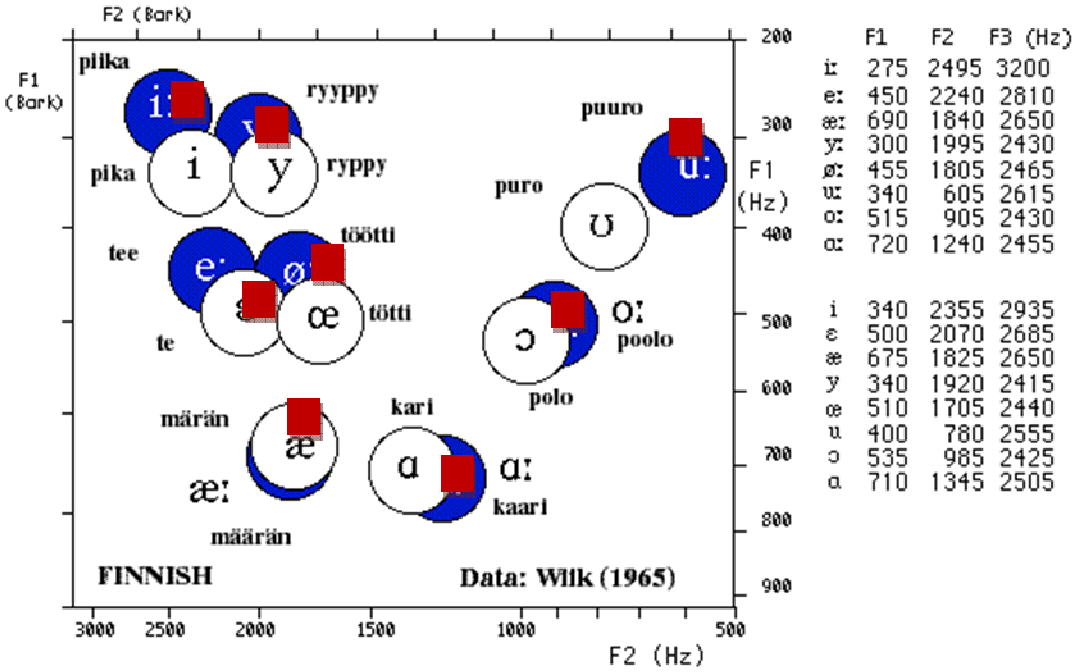


Figure 17. The frequencies of the first and the second formants obtained by the model (red squares) drawn on the Finnish vowel chart by Wiik (1965).

5.3 Varying the vocal tract shapes

After the anchor points for the variation were selected, a choice had to be made on how the variation should be performed. First, an idea of creating changes to each section individually and then going through all possible combinations was considered. However, even with only eight uniform sections this would easily lead to a vast amount of simulations. If there would be only five possible areas for each of the eight tubes around the area of the section in the anchor shape, $5^8 = 390625$ simulations would be required for each vowel. This is very time consuming and therefore highly unfeasible.

Also, many physiologically impossible shapes would occur, since the fully randomized variation would produce vocal tracts with saw-tooth shape and steep edges at locations where they do not occur physiologically. Since the main interest was in the behaviour of the first four formant frequencies, the idea of varying the vocal tract shapes using the weighting coefficients of the four corresponding cosine terms seemed reasonable.

5.3.1 Cosine transformation based profile variations

As discussed earlier, orthogonal movements of the formant frequencies by varying the corresponding four cosine series terms works ideally only when small changes are introduced to the uniform tubes. Nevertheless, the approach yields considerably good results also in other tract shapes. The standing wave patterns are changed from the original ones, but they still remain quite close to the ideal case, when the tract shapes are smooth and not very radically adjusted. Varying of cosine terms also reduces the

number of varied parameters to four, regardless of the amount of uniform sections used. In addition, one extra parameter will be later introduced for the lip rounding.

Varying the first four even cosine series terms also keeps the changes in the profile smooth. It should be reminded that by using a total of 16 sections, the tract can be completely reconstructed using 16 discrete cosine series terms, where the 16th term is a cosine whose peaks touch all the uniform tube's edges. This means that when the 16th term is not adjusted, very steep variations between two adjacent tubes will not occur.

At first, experiments were performed by adjusting directly the weighting coefficients of the DCT of the vocal tract area function. As an example, the second coefficient is adjusted in Figure 18 to constrict the tract more at the lips and to widen it at the glottis. According to the assumption of orthogonality of DCT terms in the neutral tube, this should move the first formant down in frequency. As one can see, the first formant is indeed shifted to lower frequencies, but also the second and fourth formants are shifted considerably[4].

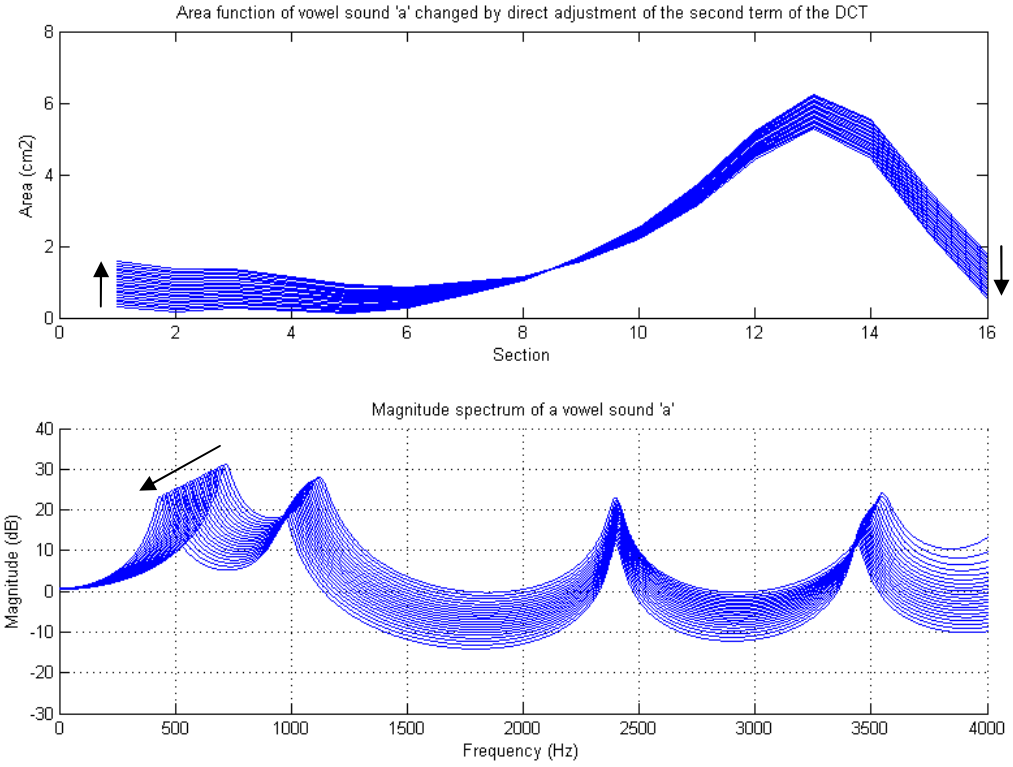


Figure 18. The second term of the DCT of the vocal tract area function adjusted. The area function is above and the corresponding spectra below. Arrows show the movement direction.

[4] The importance of the lip radiation impedance is obvious also in this picture. The first formant moves down close to one octave which should cause close to -12 dB change in the amplitudes of all the formants above the first one [**21**]. However, this is not the case because the lip opening and the radiation impedance have also changed. E.g., the amplitudes of the second and third formants have clearly changed less than this -12 dB meanwhile in the valley between them the change is about that large. Changes in the formant bandwidths have cancelled a part of the expected amplitude change.

After this experiment, a decision was made to keep the cross-sectional area of the tube at the glottal end constant when creating changes in the tract profiles. Physiologically, the area of the first section stays generally the same in similar vowel sounds and it was hypothesized that this could help in obtaining more orthogonal (independent) changes to the formant frequencies and could help to minimize the total number of tract profiles. Variation of the terms of the cosine series also varies the glottal area, so a method was needed to keep the area of the first section constant. One possibility was to scale the entire varied vocal tract up or down in a manner that keeps the area of the first tube section constant. This could be comfortably done by simply adjusting the DC-term, i.e., the first cosine series component. The result of this compensation experiment is shown in Figure 19.

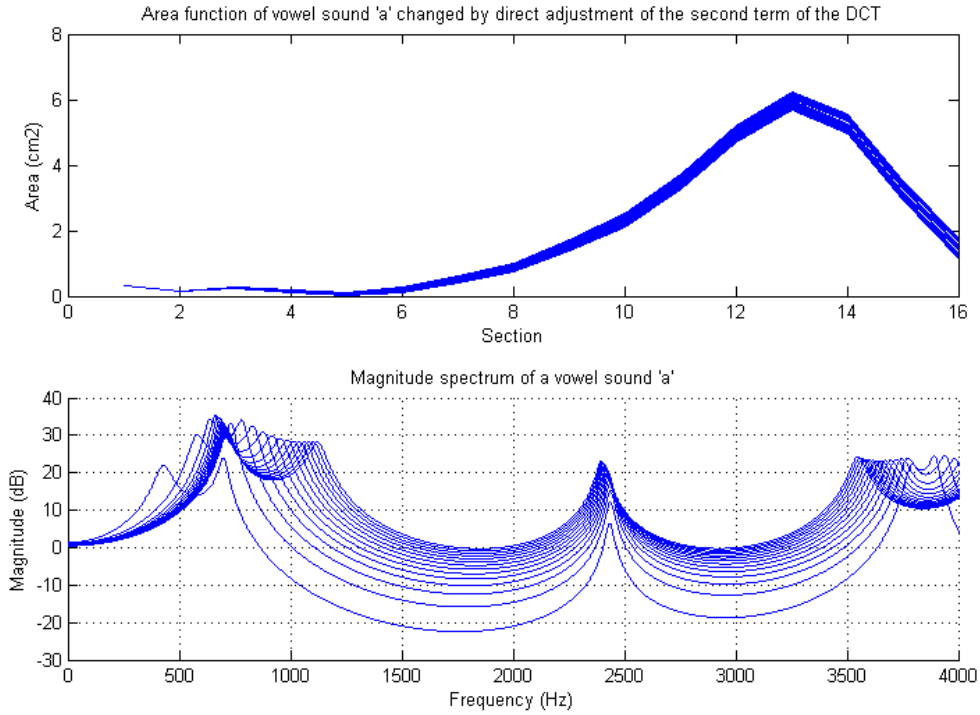


Figure 19. The second term of the DCT of the vocal tract area function adjusted. Now the glottis is held in place by compensating the movement with the DC-component.

It becomes clear from the result that solely the adjustment of the DC-term does not result into orthogonal formant changes. It can be understood by considering that when the total area function of the vocal tract is shifted up or down by a constant, the reflection coefficients are changed, because the relations between the areas of two consecutive segments change. Also this caused more easily a zero area for some junctions because the changed profile is continuously shifted down by the DC-component.

In overall, it seemed that varying the weighting coefficients of the DCT of the tract shape does not yield desired results. In theory, the orthogonality of the formant movement could be improved by finding a method of compensation in which more cosine terms are adjusted to cancel the shifting of other formants than the desired one. Such a compensation method was studied, but the task proved to be too complicated. A

more sophisticated method was required in order to solve the problem, and will be described in the next section.

5.3.2 Profile variations produced by modulation method

Adjusting merely the DC-term of the cosine transformation of the vocal tract shape causes shifting of all formant frequencies. If the original cross-sectional areas are multiplied by a constant instead, the ratios between consecutive areas stay the same, resulting in unchanged formant frequencies. This gave an idea to create a profile from the cosine series that could be used to modulate the original vocal tract shape. The produced cosine functions are called *generating functions*. Figure 20 shows an example where the area function is adjusted by the following process:

1) A vector of 16 zeros is created. The first term is adjusted to 16, so that the IDCT of this vector is a straight line with unity amplitude.
2) The second DCT coefficient of this line is adjusted to a small constant 1, causing the shape to expand at the lips and constrict at the glottis. The first value of the IDCT vector moves down.
3) The first DCT coefficient is adjusted to cancel this movement and keep the first IDCT value in unity.
4) IDCT is taken and the original tract area function is multiplied by this new generation function to obtain the first changed profile.
5) The resulting new area function is multiplied again with the same generating function in order to produce further changes (i.e., the next area function).

This method seems to give a good orthogonality property even with non-uniform vocal tract shapes. It can be seen from the Figure 20 that practically only the frequency of the first formant has been lowered. The formant frequencies can be shifted higher by using a negative constant in the DCT vector and starting the process again from the anchor tract shape. Since the modulation is always performed on the result obtained in the previous phase, the increase in constriction occurs in a logarithmic manner and zero areas are almost never reached.

The cosine functions used to move primarily the four first formants are shown in Figure 21. Dashed line shows the smooth sensitivity function of 300 points and solid line the corresponding sensitivity function using only 16 DCT-terms. It has to be noted that the discrete sensitivity functions are symmetric about the center point of the vocal tract. In the figure it means symmetry about the point (150, 0). This preserves the orthogonality even in the discrete case, because the inner product between any two differing discrete shapes stays zero.

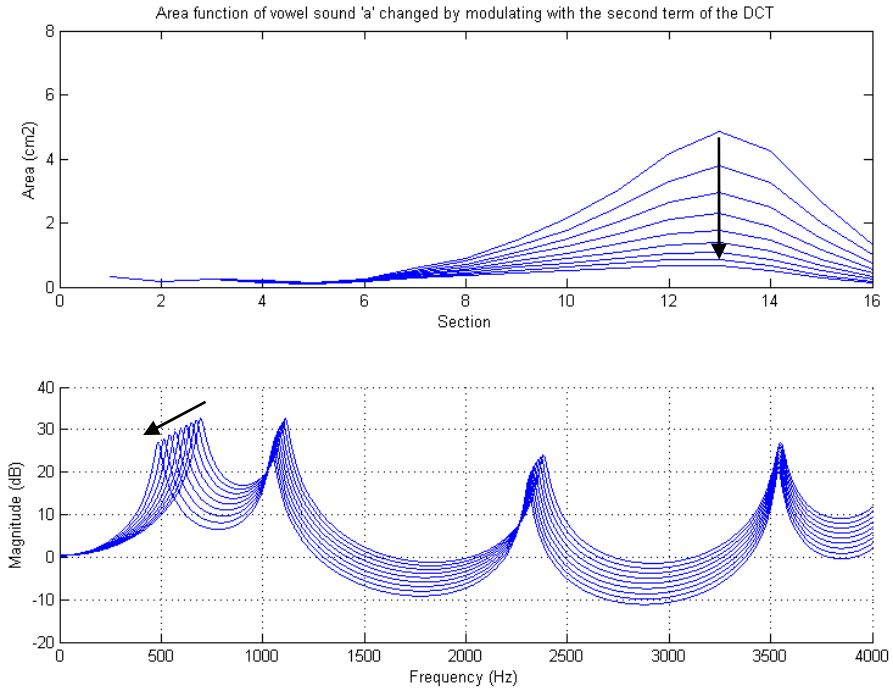


Figure 20. Vocal tract area function adjusted by creating a profile by changing the first cosine term and using it stepwise in modulating the original area function. Glottis is held in place by adjusting the DC-component in the alteration profile.

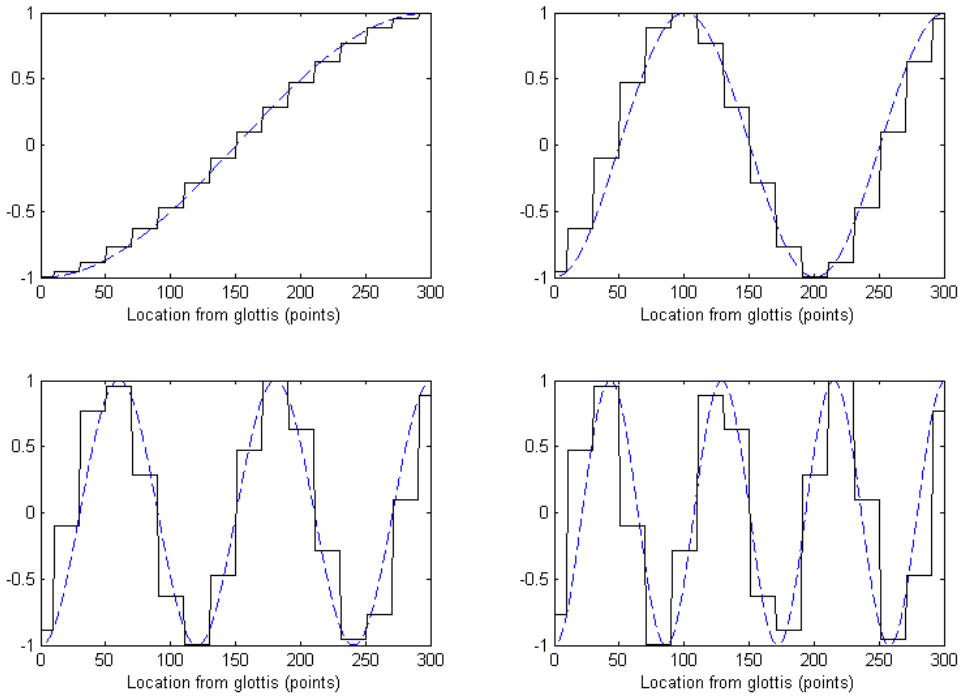


Figure 21. The four first cosine functions of even order applied for the generating functions. The smooth sensitivity function is shown with dashed line and the corresponding 16-length profile shape with solid line.

The modulation method described above was applied in this study as the main tool to create variations in the vocal tract shapes. The transition of the formant frequencies is not completely orthogonal due to the fact that the vocal tract is not uniform and the standing wave pattern is thus changed from the ideal case. Depending on the vocal tract shape the orthogonality is sometimes preserved better and sometimes worse. The closer to a uniform tract the anchor shape is the better is the result in orthogonal formant shifts. In any case, this method reduces the amount of parameters in the variation process into four plus the lip length parameter, when compared to the variation of every section individually.

Figure 22 shows how variations are made to the anchor vocal tract shape of the Finnish vowel /a/ using the profiles corresponding to the four first formant frequencies. 5 iterations towards more constricted and 5 iterations towards less constricted tract shapes are made. The corresponding changes in the spectrum are shown on the figure on the right. The orthogonality suffers as more variation is introduced. Also, the orthogonality is reduced by the use of lip radiation impedance. The lip radiation is taken into account in the figures.

In the case of vowel sound /a/, the changes are not as orthogonal as they would be for the uniform tube, but it can still be quite well predicted which formants are likely to move with which coefficient. The first formant is shifting considerably less than the other formants. The overall shape of tract is maintained, because the variation is relatively much smaller in the parts where the area is small in the original shape. It is characteristic for the /a/ to be more constricted at the glottis and the tract opens up when approaching the front part of the mouth cavity.

In the appendix similar figures with captions illustrate, how variations are made to the vocal tract shapes of Finnish vowel /i/, and also to a uniform tube. The uniform tube is discussed in connection to Figure A-1, and vowel /i/ in connection to Figure A-2.

The described modulation method is used in filling in the formant space. In the ideal case, it gives nearly orthogonal formant shifts by varying the vocal tract area functions, keeping them still reasonably close to the original anchor shapes and therefore maintaining at least some of the articulatory plausibility.

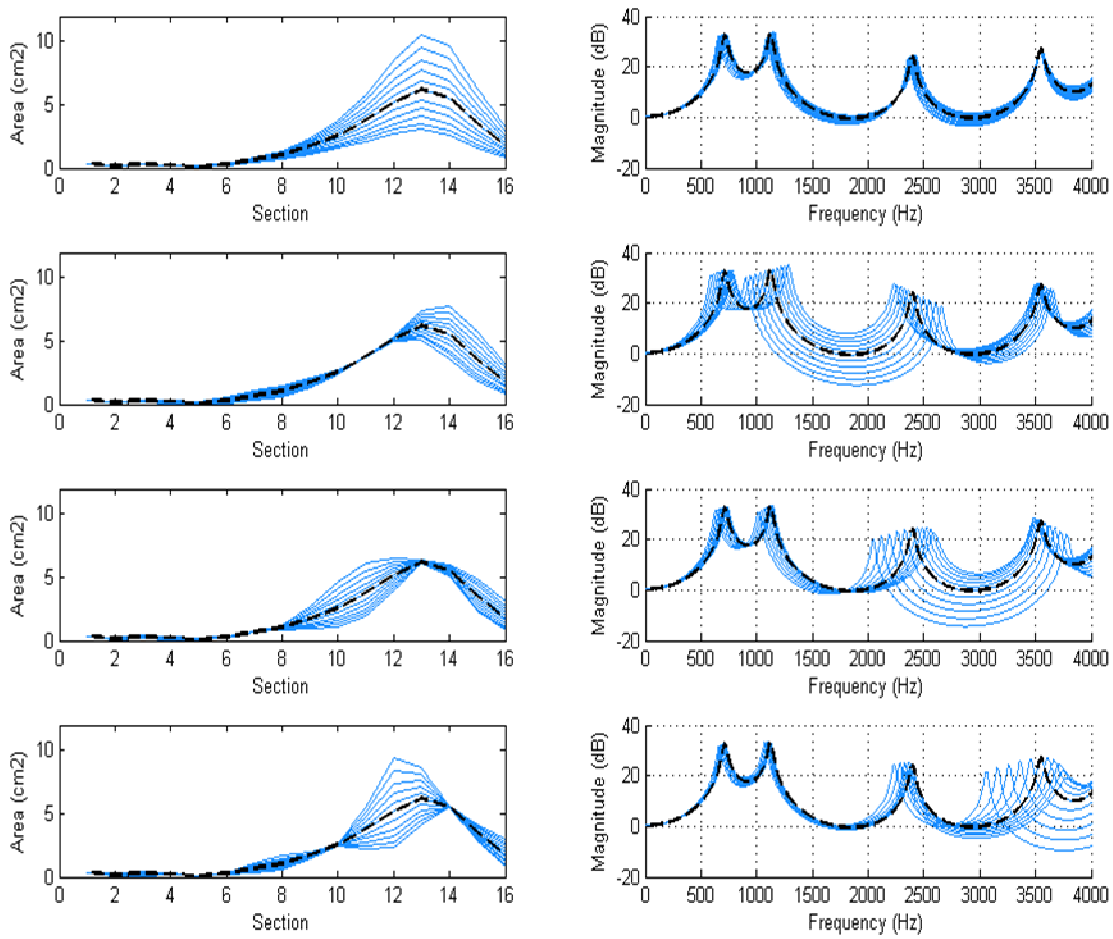


Figure 22. Variations created to the area function of the Finnish vowel sound /a/.

5.4 Filling in the formant space

In order to obtain a wide variety of different speech sounds, many different vocal tract shapes are simulated and the formant frequencies for each shape are extracted. The simulation is performed with the modulation method described in section 5.3.2. All the perturbation combinations related to the four first formants are simulated by creating vocal tract changes in a stepwise manner:

1) A vector of 16 zeros is created, and the first term is given the value 16 to obtain a straight line with unity amplitude. IDCT of the vector is used in modulating to obtain the original anchor tract shape.
2) The 8th cosine coefficient (related to function of order seven) is changed to one, and the DC-term is adjusted so that the cross-sectional area of the tube at the glottal end stays constant. The generating function is created by taking the IDCT of this vector.
3) The original area function is multiplied by the generating function. The areas obtained are fed into the vocal tract model and the obtained formant frequency data is stored.

4) The resulting area function is again multiplied by the generating function at each step until all iterations to lower the fourth formant are done. All data is stored.
5) Generating function to raise the fourth formant is created by changing the 8th cosine coefficient to -1 and adjusting DC term correspondingly. Similar iterations are done starting again from the original area function with the new profile.
6) After all iterations for the fourth formant, a new profile is created for the third formant by changing the 6th cosine coefficient similarly as in step 2.
7) After every individual change in the third formant all the iterations for the fourth formant are repeated.
8) All combinations are gone through by changing gradually also the second and first formants.

When the original area function has a high value at some location, the modulation can cause very high expansion to the vocal tract shape. Also, with a high degree of constrictive modulation, the area function may be very close to zero at some location. In order to maintain the physiological correspondence of the vocal tract it has to be assured that the area function of the resulting vocal tract does not reach zero at any location. It is also equally important to avoid too big areas. In order to maintain realistic tract shapes, the section area range is limited between 0.03 cm^2 and 12 cm^2. 12 cm^2 is chosen for upper limit because it is at the limit that human vocal tract can reach according to the vocal tract area functions appeared in the literature. 0.03 cm^2 is chosen as a small cross-sectional area that still provides cues for consonants or consonantal transitions without compromising the functionality of the model.

The number of iterations and the weighting coefficient were chosen manually based on simulation experiments with a small amount of samples. If the weighting coefficient or the number of iterations is chosen to be too small, only small clusters of points are obtained around the anchor points. When the modulation values are increased until the formant clouds begin to overlap across vowel categories, a large variety of vocal tract shapes is achieved. The first and second formants are varied with a larger number of iterations since they affect the main characteristics of the vowel sound. The third and the fourth formants contribute mainly to the timbre. In total, six expansive and six constrictive iterations were used for the first and second formants, three expansive and three constrictive iterations for the third formant, and only one expansive and one constrictive iteration for the fourth formant. Four discrete values for the lip rounding were given also to give more variety to the modeled sounds. These values were chosen uniformly between the minimum length used in vowel sound /i/ (0.1 cm) and the maximum used in anchor point of vowel sound /u/ (1.1 cm). As could already be seen in Figure 14, lengthening of the lips affects considerably the first and second formants. This caused further spreading of the clouds around the starting point.

Using the above choices for parameters, a total of $13 \times 13 \times 7 \times 5 \times 4 \times 9 = 212940$ iterations were done. The values 13, 7 and 5 result from the expansions, constrictions plus the original shape, 4 from the lip rounding, and 9 from the simulation of all 8 vowels plus the uniform vocal tract. During these simulations, 9343 tract areas exceeded the limit of 12 cm^2 and the corresponding section areas were limited to this value. None of the shapes reached zero area. All simulated vocal tract shapes and

corresponding formant frequencies were stored into a lookup table ***T*** for further processing.

All the simulated combinations between the variations of cosine terms and lip rounding for all vowel sounds are drawn in an F1-F2-plot in Figure 23. The plot shows all simulated points according to their first and second formant frequencies In addition, a *Voronoi diagram* of Finnish vowel categories is superposed to the image, allowing estimation of the vowel category boundaries with respect to the simulated variations[5]. The diagram is drawn in every F1-F2-plot from this on for illustrative reasons.

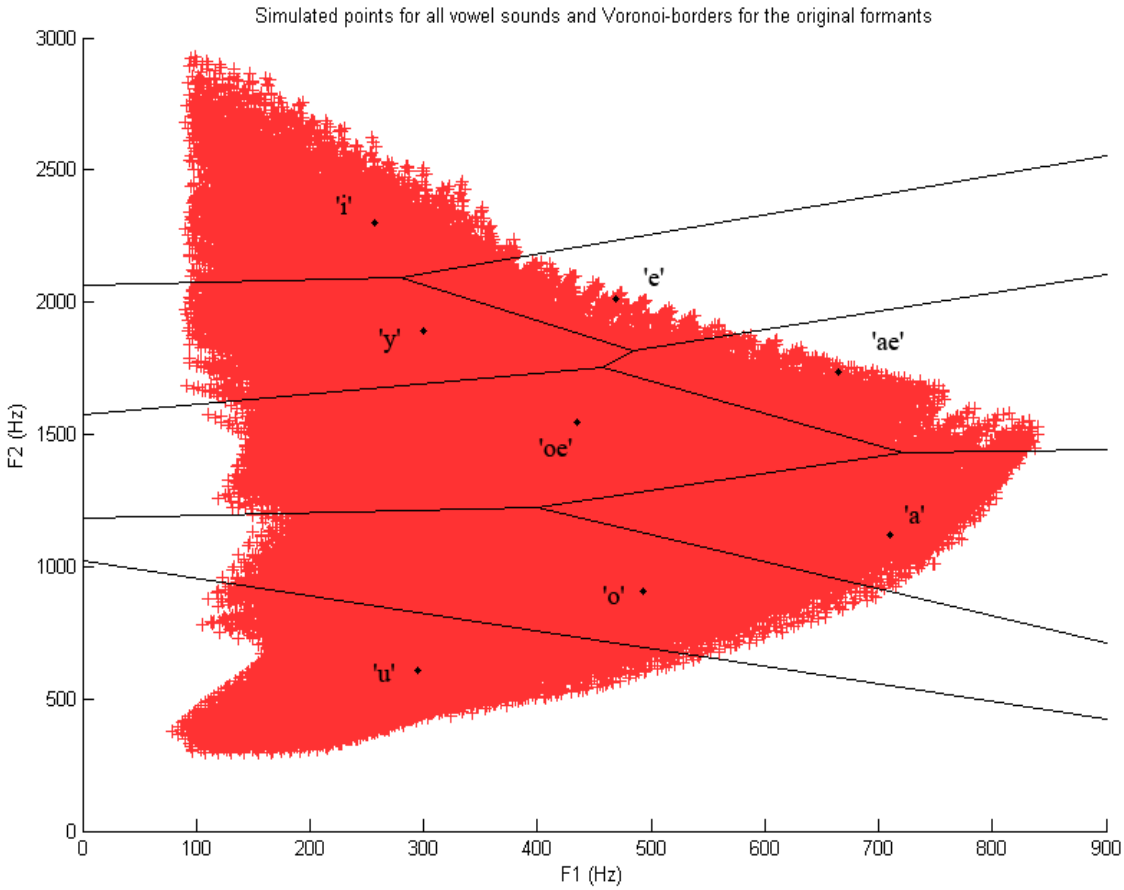


Figure 23. The simulated formant frequencies of the vowel variations plotted in F1-F2 space. The labeled points show the original places of the first and second formants for each vowel category. The lines are Voronoi-borders that describe boundaries between vowel categories in terms of Euclidean distance.

The density of the simulated points for the eight Finnish vowel sounds in the F1-F2 plane is illustrated in Figure 24. It can be noticed that the first formant has a maximum value somewhere above 800 Hz and a minimum value at approximately 100 Hz. The second formant varies in a larger scale. In Carré's reseach, congruent results were obtained [18]. Carré states that the first formant rapidly reaches asymptotic limits outside the range of 200-800 Hz.

[5] The linear Voronoi-borders for vowel categories may differ from the perceptual limits observed by a listener. Voronoi-borders are only used to give approximate information about vowel categories.

The low density line in the middle of the figure divides the vowels into back vowels /a/, /o/, and /u/, and the front vowels /i/, /y/, /e/, /ae/ and /oe/. The vocal tract shapes do not easily cross this limit during the variation process since it would require a strong constriction at one end of the tract and a strong expansion at the other, which is not characteristic to the used modulation procedure. Some points at this low density region are still simulated, and they provide information regarding the transitions between front and back vowels. The density is highest in the vicinity of points (300 Hz, 500 Hz) and (500 Hz, 1500 Hz). The former high density area is explained by the clouds formed by vowel sound variations for /a/ and /o/. They both mainly spread down to the region of vowel sound /u/. The latter high density region is explained by the fact that all the back vowels spread towards this particular area. It can also be noted that the region at (500 Hz, 1500 Hz) corresponds to the first two formant frequencies of a uniform tube. The separate spreading regions for each vowel category can be observed in the appendix B (Figure B-1, Figure B-2, Figure B-3 and Figure B-4).

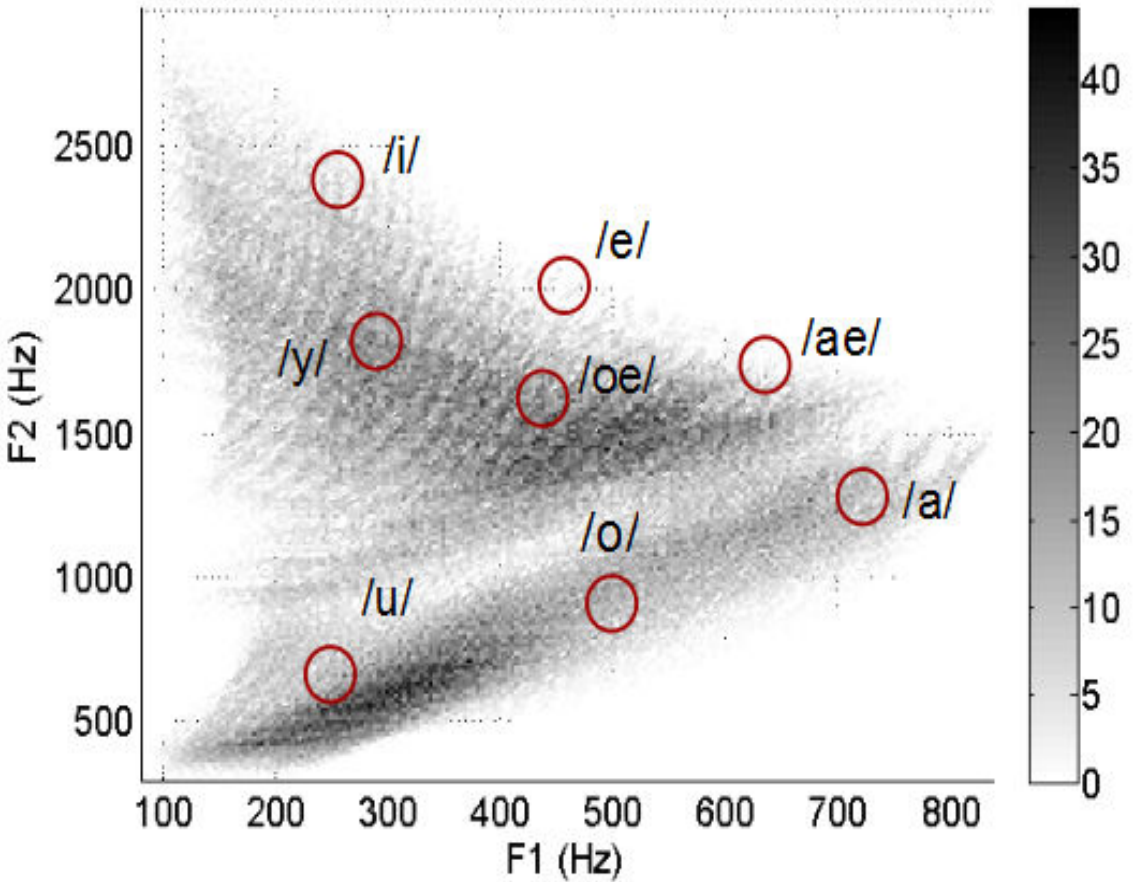


Figure 24. Density of the simulated points in F1-F2-plane. The dark areas have a higher density of points. In total 189280 variations corresponding to all eight Finnish vowel sounds are plotted. The formant frequencies for the anchor shapes are indicated with circles.

For comparison, a similar density plot of formant frequencies for a Finnish person speaking has been plotted in Figure 25. Formants have been tracked for 52951 time frames of speech. It can be seen that the size of the vowel triangle is somewhat smaller than in the simulated case, but the low density region is still visible. The size of the vowel triangle, i.e., the locations of the first two formant frequencies depends on the vocal tract length, speaker-specific way of pronunciation. Also coarticulation reduces formant frequencies at the extremities of the triangle.

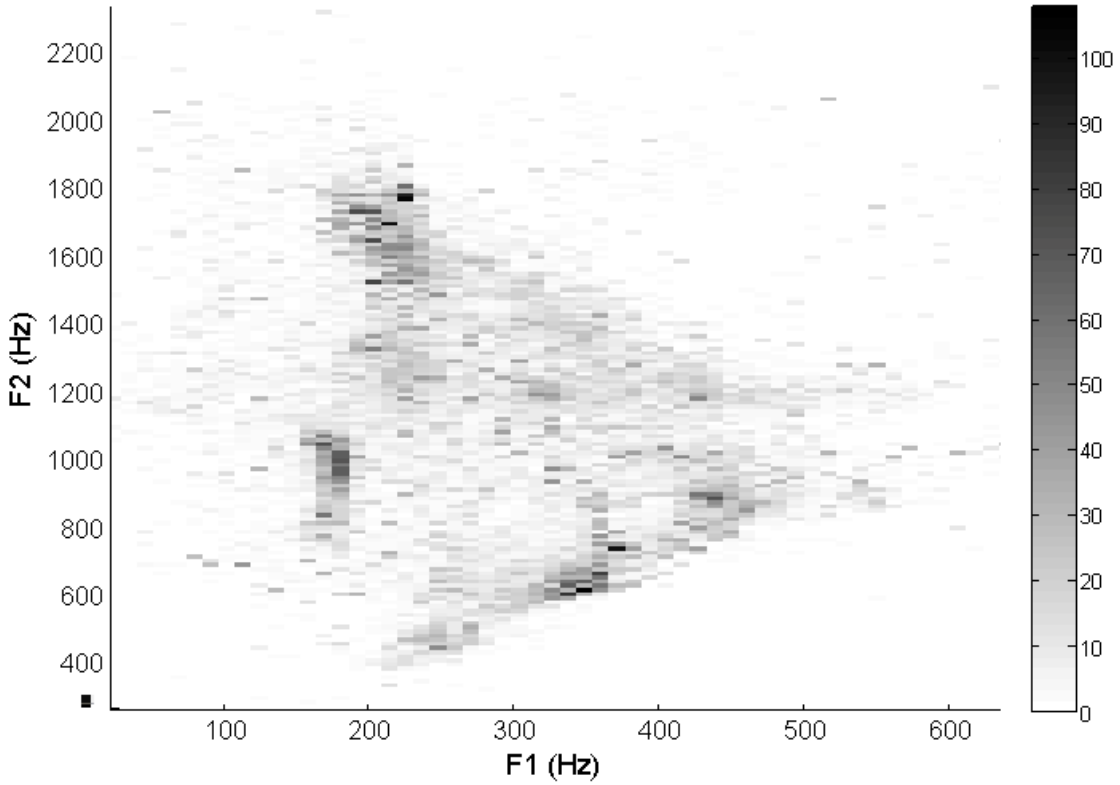


Figure 25. A density plot for formants of 52951 time frames tracked from speech of a Finnish person.

6 Estimation of instantaneous and dynamic vocal tract shapes

This section describes how the lookup table can be used in estimation of area functions for spoken vowel sounds. First, a short description is given on how the formant frequencies from speech are extracted, followed by a method for finding a smooth path through a selection of area functions found for each time window. Finally, waterfall plots illustrate the found vocal tract shapes through the speech signals.

6.1 Formant analysis of speech signals

Linear predictive analysis is used for extracting formant frequencies from recorded speech signals as explained in section 3.6.1. The test signals were recorded using a normal computer microphone with 16 kHz sampling frequency. The prediction order was chosen to be 22, since smaller order models had occasional problems in modeling two formant frequencies situating close to each other. This was especially a problem in the case of vowel /u/, where the two actual first formants were detected as one formant and the third formant was detected as the second one. Before the linear prediction, the speech signal was highpass filtered with a first order *pre-emphasis filter* of the form

$$P(z) = 1 - \mu z^{-1}. \tag{47}$$

This removes a typical spectral tilt of -6 dB/octave, producing a similar flat spectrum as seen in the vocal tract simulations (volume velocity transfer function). After pre-emphasis the formant frequency estimation works better and the comparison to the simulation results is more reliable. The coefficient μ for this filter can be chosen between 0.9 and 1 with little significance [37], and for this work it was chosen to be 0.9.

Linear predictive analysis is performed for the entire speech signal using a Hann window of 25 ms in length, separated by 10 ms intervals. The four first formant frequencies for each time frame are stored into a matrix. If the absolute value of a root inside the unit circle is more than the chosen threshold value, 0.9, the corresponding root is chosen as a formant. If less than four formants are found inside a frame, the sample of the frame is considered as unvoiced, and formant frequencies are not stored. Figure 26 shows the detected formants in a continuous sound proceeding through the vowels /a/, /u/, and /i/. As can be seen, there are no radically deviating formant frequencies outside the expected regions. Transitions between vowels can be seen as paths from one formant cloud to the other.

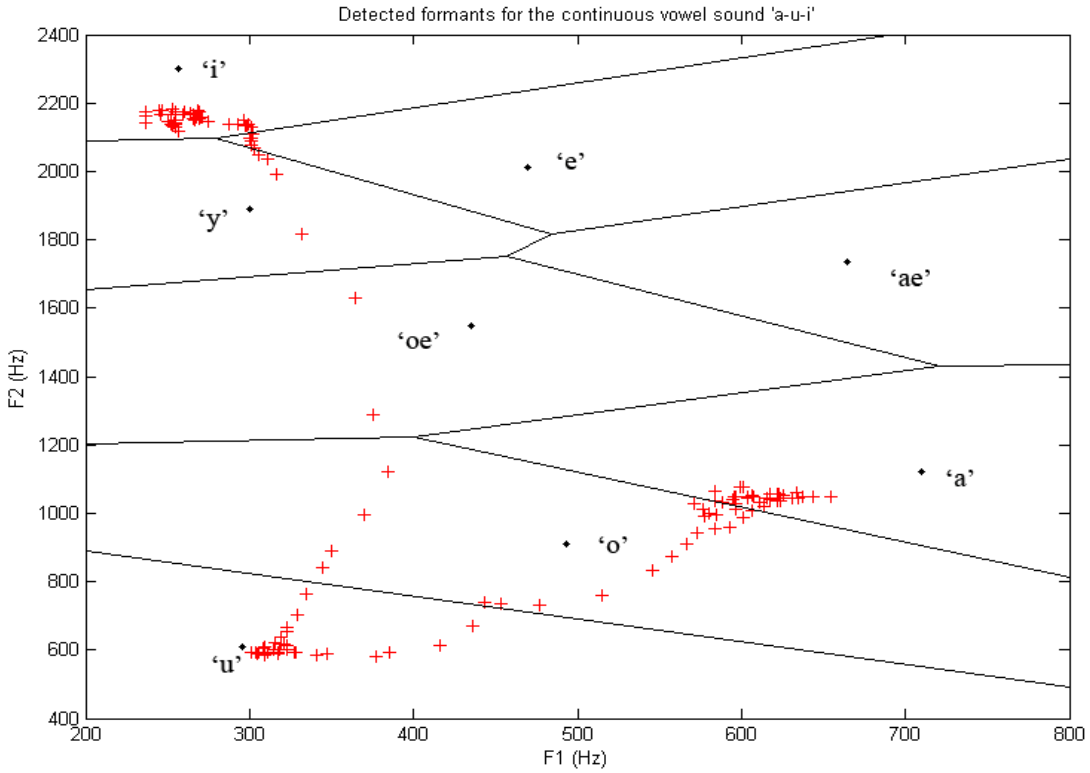


Figure 26. Detected first two formants for a recorded transition through vowel sounds /a/, /u/ and /i/.

6.2 Estimation of instantaneous VT shapes from formant frequencies

The aim of the entire work is to find correspondences between the detected formant frequencies and the modeled vocal tract shapes. The profile candidates are selected from the lookup table based on their formant frequency deviations from those estimated from the actual speech frame. The allowed deviation is ±60 Hz in this work. The value of this parameter is not crucial, since later only the best matches from the formant group inside the borders are included in further search. Nevertheless, the limit should be wide enough to provide a good variety of vocal tract shapes for each time window for the forthcoming trajectory estimation algorithm.

At first, the entire lookup table ***T*** (see section 5.4) is searched for points where the first formant is located inside the given limits. This subgroup is again searched for the similar correspondence for the second formant, and in the third phase a subgroup is formed for the third formant. The fourth formant is left out from this grouping since the subgroup matching all three formants resulted in convenient group sizes from tens to a few hundred points per frame. The subgroups matching all four formants were observed to be considerably smaller and sometimes empty.

Now, an error value for each element inside the group is calculated. The error is simply the sum of the differences between all the three formant frequencies[6]:

$$e_i = \left|F_1 - F_{1,i}\right| + \left|F_2 - F_{2,i}\right| + \left|F_3 - F_{3,i}\right| \qquad (48)$$

where F_k is the k:th formant frequency of the actual speech frame and $F_{k,i}$ the k:th formant frequency for the candidate i. From the group, a certain amount of minimum error values are selected for each frame, and the vocal tract area functions corresponding to them are used as competitors in finding the best dynamic vocal tract shape.

It is important to have several competitors for each time frame, because it may occasionally be the case that the original formants are tracked inaccurately, or too small a frequency limit has been used in candidate selection, leading to a small set of profiles where none of the members can be considered as a fluent articulatory continuation of the previous one. This may lead to a sudden change in the vocal tract shape trajectory distracting the path optimization algorithm. In order to avoid these problems, the minimum group size was limited to 5. If smaller groups are found, the time window corresponding to them is discarded. The maximum amount of competing vocal tract shapes was chosen to be 50 in this work.

6.3 Estimating dynamic VT shape

Ensuring a smooth and in an articulatory manner plausible change of the vocal tract shape along a dynamic speech signal is an important but also complicated task. Without any smoothing algorithm, the *many-to-one* characteristic of the mapping would become a considerable problem. Many vocal tract shapes may have the same spectral characteristics and choosing strictly the best match according to the first three formants at every time instant would cause abrupt changes in the VT shapes, as will be later observed.

The basic presumption for the vocal tract movements is that the tract shape can change only a small amount between two consecutive time frames. Also, it can be observed that the articulatory movements during normal speech tend to minimize the usage of muscular energy. Observing the formant transitions in continuous speech shows that the vowels often blend together and lose some of their characteristics that would be observable in vowels produced in isolation. The effect is well known in the phonetic theory and it is called *co-articulation*. This creates a plurality of allophones for every phoneme of a language.

Using these assumptions, a simple trajectory estimation algorithm was created that looks after the shortest (and smoothest) path through the space of possible VT shapes.

[6] In later experiences while writing a pending conference paper on the subject, psychoacoustically more reasonable choices were made for the frequency deviations as well as weightings when calculating the error. Deviations of 20 Hz, 60 Hz and 200 Hz in formant order were used, and the differences in the lower formant frequencies were weighted more in (48). This is due to the fact that auditory frequency resolution is higher around the lower formants.

The 5-50 best candidates for the vocal tract area function at each time window are collected and set for further processing.

The path minimization problem follows the Bellman's optimality principle, stating that if an optimum path from point *A* to point *C* goes through point *B*, also the paths from *A* to *B* and *B* to *C* are optimal paths [38]. Because the path will be minimized across each time window, iterative processing can be used to find the global optimal path. First, all pairwise distances between all possible consecutive vocal tract shapes in the candidate sets, $e_{ij}(S_i(t), S_j(t+1))$, are calculated using Euclidean distance. The lip length is added as the 17th element in the area function vectors. If there are *T* time windows in the signal, *T-1* matrices are obtained containing the Euclidean distances. The Euclidean distance matrix for each time frame is of the form

$$\mathbf{E}_t = \begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1N} \\ e_{21} & e_{22} & \cdots & e_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ e_{N1} & e_{N2} & \cdots & e_{NN} \end{bmatrix} \tag{49}$$

Where e_{ij} indicates the distance from the node *i* of time instant *t* to the node *j* of time instant *t+1*.

Next step is to find the target nodes where the state of the system is likely to shift from each node of each time instant *t*. These nodes are simply the indexes *j* of the minima of each row in the Euclidean distance matrix. Now following these indexes continuously, the shortest path through all time windows is searched starting from each node at time instant 1. This leads to *P* different paths, where *P* is the amount of competitors at the first time instant. The winner provides the minimum total distance, i.e., is chosen as the globally shortest shape trajectory.

The method is illustrated in Figure 27. Integer numbers are used to denote different vocal tract shapes. Now the Euclidean distance between nodes is simply the difference between the two numbers, and the arrows show the shortest path through the system starting from every possible state of time 1. Maximum of 6 competing states are used, and X marks the nodes where not enough states were found. In this example the total distances for the five paths are (3, 1, 7, 2, 4) starting from the upmost one at the first time instant. The shortest path is thus chosen to be the path 2, starting from node 2. In the actual system it is rarely the case that the distance would be the same to two different nodes. If such a situation occurs, the first of the values is chosen.

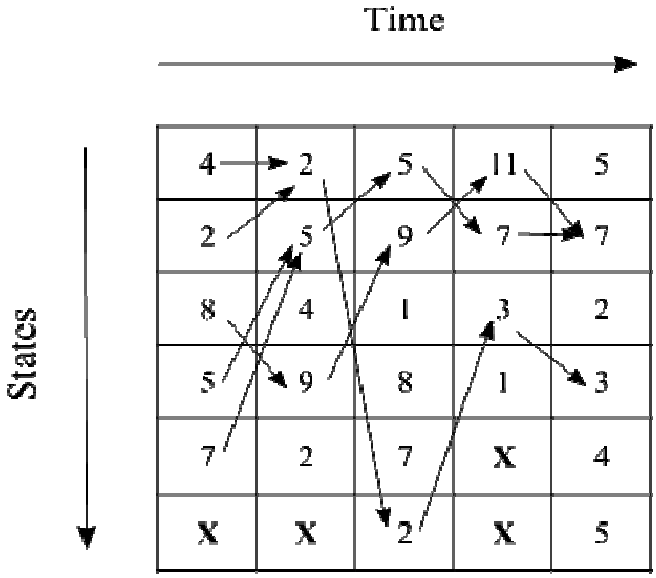


Figure 27. Shortest paths through a system starting from all the possible states at time 1.

This method was tested and it was observed to give good results for continuous transitions between vowels. The simplicity of this method has also downsides. For example, if the tracking is started at a bad location during the speech signal, the candidates for the first state may all be bad. This may lead to a situation where, e.g., formant frequencies corresponding to a vowel sound /a/ are found but the closest vocal tract shapes may have the maximum expansion of 12 cm^2 at the front part of the mouth. Such a big expansion at the first frame could then lead to a series of vocal tract shapes having maximally expanded shapes throughout the signal, unless a frame with only smaller corresponding area values is found and all the area functions are forced to drop down. Use of a larger number of competing states reduces the problem. Also, making more restrictions to the articulatory-to-acoustic mapping would reduce these extreme vocal tract shapes that are physiologically possible, but occur only rarely in speech.

If this minimization method was to be used for speech signals including also unvoiced segments, the method should be improved. Otherwise problems will occur especially in situations where there are segments without reliable formant information. During these segments, the actual vocal tract may shift to a completely different shape, and when the formants can be tracked again, the movement minimization process should be started again from all the area functions corresponding to these new formant frequencies. For example in words like *"esa"* ([esa]), the minimum movement assumption may get lost during the fricative [s]. More complicated path optimization methods are not experimented in this thesis, but research concerning them would be an essential part of future work that aims at improving the vocal tract area function tracking process.

6.4 Examples of VT shape trajectory estimation from continuous speech

A vocal tract shape trajectory estimation test was made for several self-recorded speech signals. First, the tracking method was used for the vowel triphtong "/a/-/u/-/i/" whose detected formant frequencies were already drawn in Figure 26. The detected vocal tract shapes are drawn in a waterfall plot, where time is on the x-axis, vocal tract section number on the y-axis, and the area on z-axis. The waterfall plot is drawn in Figure 28, and the formant frequencies for the detected final vocal tract shapes in Figure 29. Lines connecting two consecutive formant points are also drawn, giving trajectory information of the formant movements. It can be seen that at time 0, the area function is as expected for the back vowel /a/; constricted from the section close to the glottis and expanded at the front part of the oral cavity. The area at the maximum extension is at about 4 cm^2.

Moving forwards in time, the vocal tract starts constricting from the front part and at the time of 0.5 s approximately, the lip section is longer than previously due to the lip rounding of vowel /u/. During /u/, the vocal tract shape does not remind exactly that of the original anchor shape due to the fact that the same formant frequencies are reached when the front half of the vocal tract is constricted, as when the first half closer to glottis is expanded. It still remains a question what really happens in the vocal tract in this kind of a transition. Because the similar vowel sound /u/ can be created also with lowering the front part of mouth, it is possible that the vocal tract prefers to move to /i/ through this state, if this kind of transition needs the minimum amount of energy. As can be seen in Figure 29, the first two frequencies still correspond to the vowel sound /u/ with the area function at around time instant 0.7 ms.

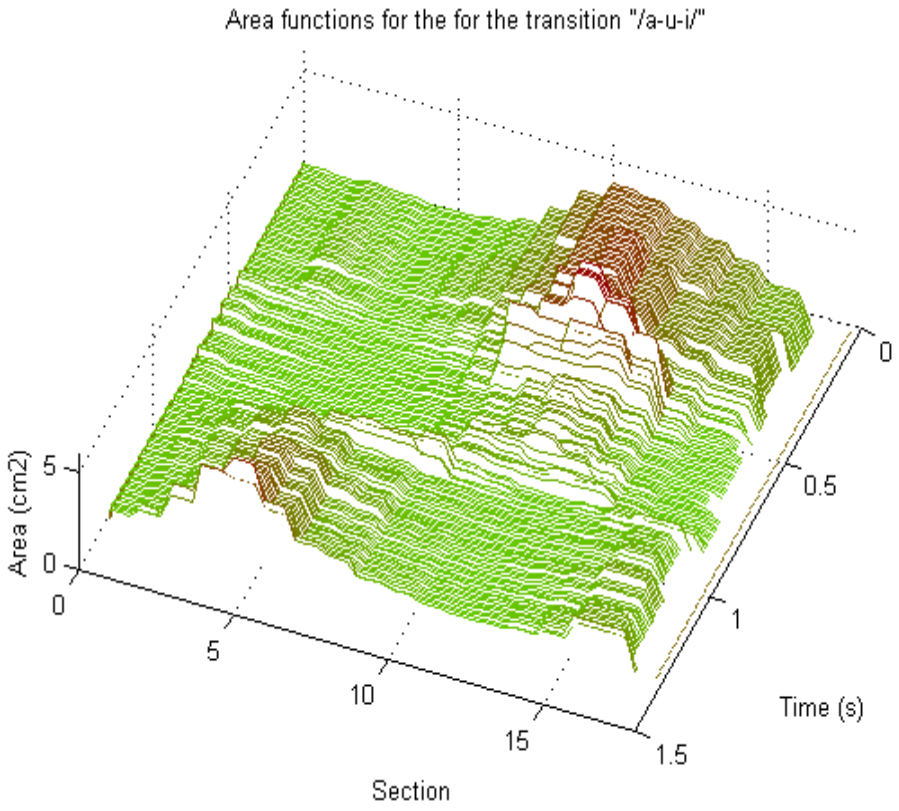


Figure 28. Waterfall plot of the area functions of a continuous vowel transition "/a-u-i/".

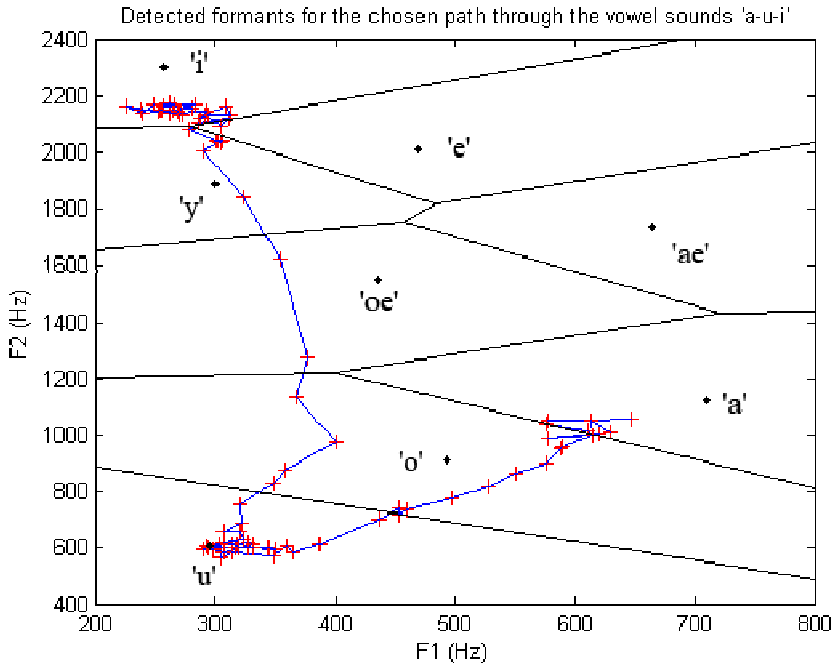


Figure 29. Formant frequencies corresponding to the final detected vocal tract shapes of the transition ”/a-u-i/” (plus signs) with the trajectory through the formant space (solid line).

When the vowel sound /i/ is approached, the lips are shortened again and the vocal tract shape starts to remind accurately the anchor shape. During the transition from vowel sound /u/ to /i/, it is seen how the lip section and the first half of the tract start gradually growing in area, while the expansion characteristic to back vowels around section 11 starts constricting down. It is also interesting to notice from the formant chart, that the transition between vowel sound /u/ and /i/ does not form a straight line but tends to bend more towards the neutral tract shape.

For comparison, the waterfall plot for the vowel transition /a-u-i/ was also plotted without the optimization algorithm in Figure 30. Without optimization, the closest match to the detected formant frequencies is always used regardless of the adjacent windows. As can be observed, the shapes jump up and down in different windows with no articulatory reliability. A large number of shapes reaching the 12 cm^2 limit are detected. It becomes clear that the optimization algorithm is a vital part in the detection of vocal tract shape trajectories.

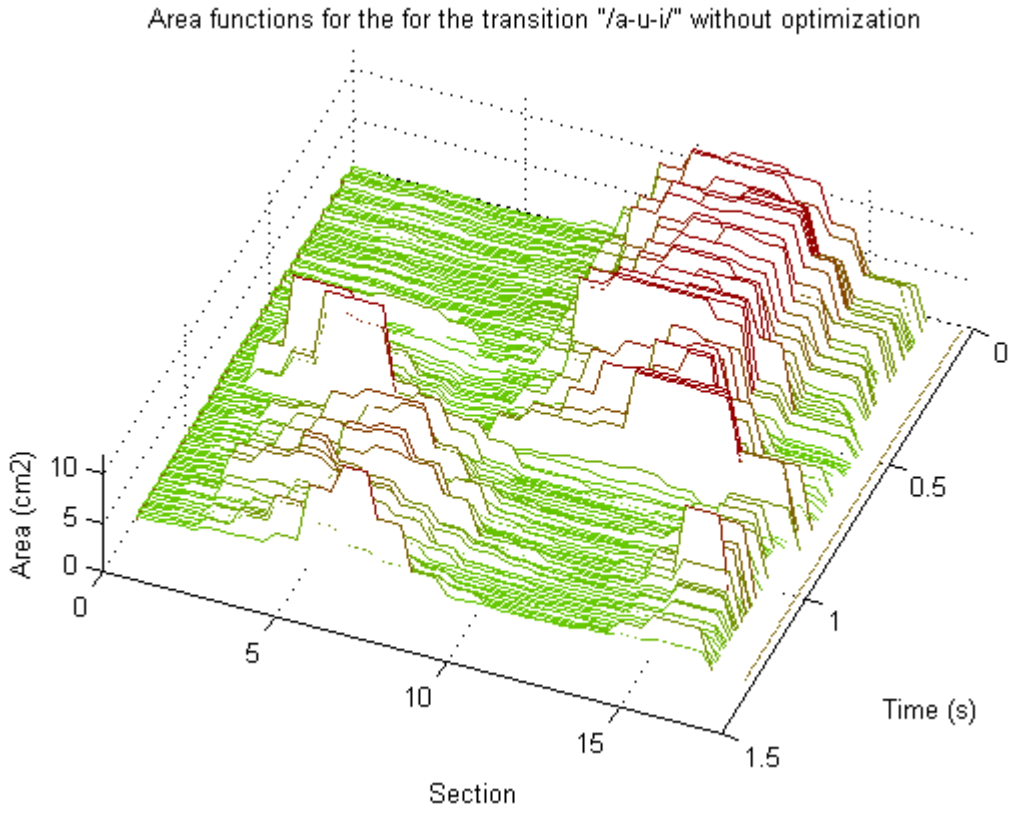


Figure 30. Waterfall plot of the area functions of a continuous vowel transition "/a-u-i/" without the optimization algorithm.

The optimization test was also performed for the transition through the vowel chain "/y-e-ae/". The waterfall plot can be seen in Figure 31 and the formant information in Figure 32. Again, the vocal tract shape starts from a shape that reminds the anchor shape for /y/ but is scaled down in area. As discussed earlier, scaling of areas with a constant factor does not affect the formant frequencies of the KL-model. The lip rounding is modeled in an expected manner for /y/. Soon afterwards, the vocal tract expands to remind the original anchor shape of /y/, and from there towards the accurate vocal tract shape of vowel sound /e/. From there the transition continues smoothly to the final tract shape of vowel sound /ae/.

In the next test, the tracking for the word "Anna" ([an:a]) was examined. The resulting waterfall and formant plots are shown in Figure 33 and Figure 34. The word contains a long nasal sound with formant frequencies F1 and F2 around 270Hz and 1200 Hz respectively. The vocal tract shape detector considers this sound as an almost closed area function, as it should, considering the fact that the oral cavity is closed during pronunciation of nasal [n]. An important observation in this word is the transition from /a/ towards the nasal, which introduces a gradual closing of the area around the sections 15 and 16. This can be interpreted as an articulatory process where the tongue approaches alveolar ridge in order to close the vocal tract at the front part of the mouth. Thus the method seems to illuminate interesting details even related to sounds and sound combinations which are not explicitly modeled with the present KL-variant. Remember that the model does not have a nasal tract yet.

One area function tracking experiment was performed with a different male speaker saying the word "yhdeksän" ([y][h][d][e][k][s][ae][n]) using stable speech tempo. The duration of the entire word is around one second. The waterfall plot and formant plot can be seen in Figure 35 and Figure 36. It should be noted that because formant structure of some sounds could not be detected, a number of frames are skipped during the sound. The length of the detected part is thus around 0.6 s.

As the waterfall plot in figure Figure 35 illustrates, the vowel parts of the signal are detected rather well. From 0 to 0.15 s, the area function reminds of the original shape for vowel sound /y/. The area function for /e/ is detected around 0.3 to 0.4 s. Finally, the vocal tract shape for /ae/ is located around the time of 0.45 s. In the end, the transition to /n/ is again modeled with the front part constriction, and the formant frequencies lie around the same point as in the previous word "Anna". It can be seen in the corresponding formant plot that the formants never reach the characteristic region of /e/. This is due to the effect of coarticulation and the principle of energy minimization in vocal tract movements. Vowel sounds in rapid speech are not pronounced as clearly as they would be in slow speech or when pronounced separately. All the vowel sounds are partly blended together and this causes the tract shapes of /e/ and /ae/ also to remind more of the mixture of both.

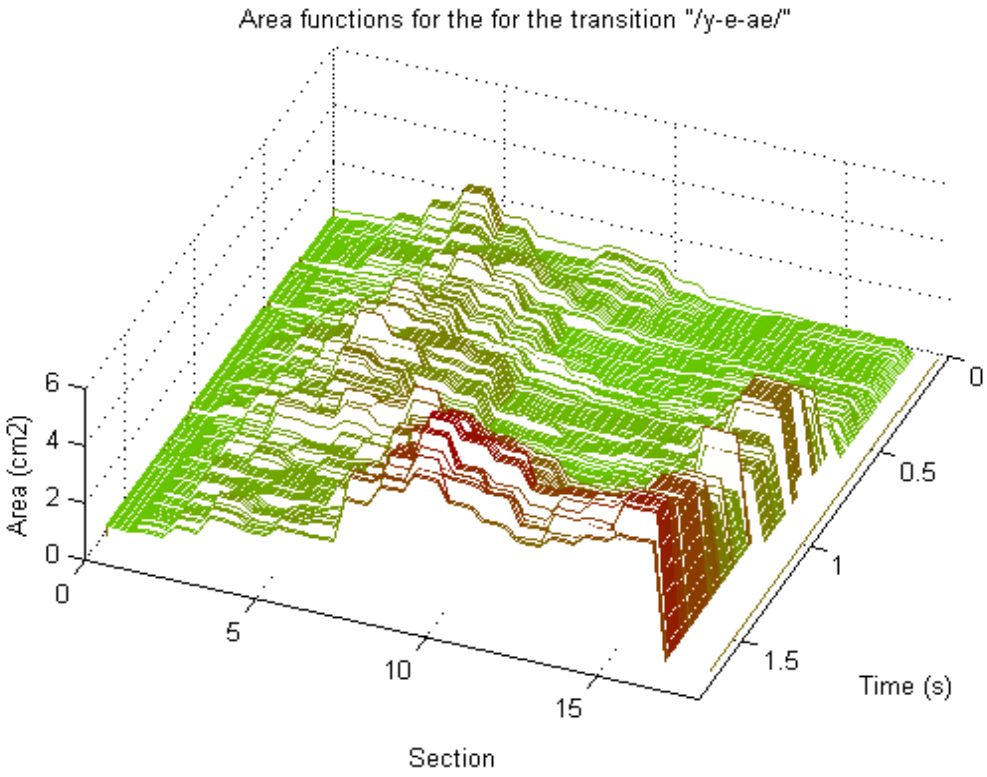


Figure 31. Waterfall plot of the area functions of a continuous vowel chain "/y-e-ae/".

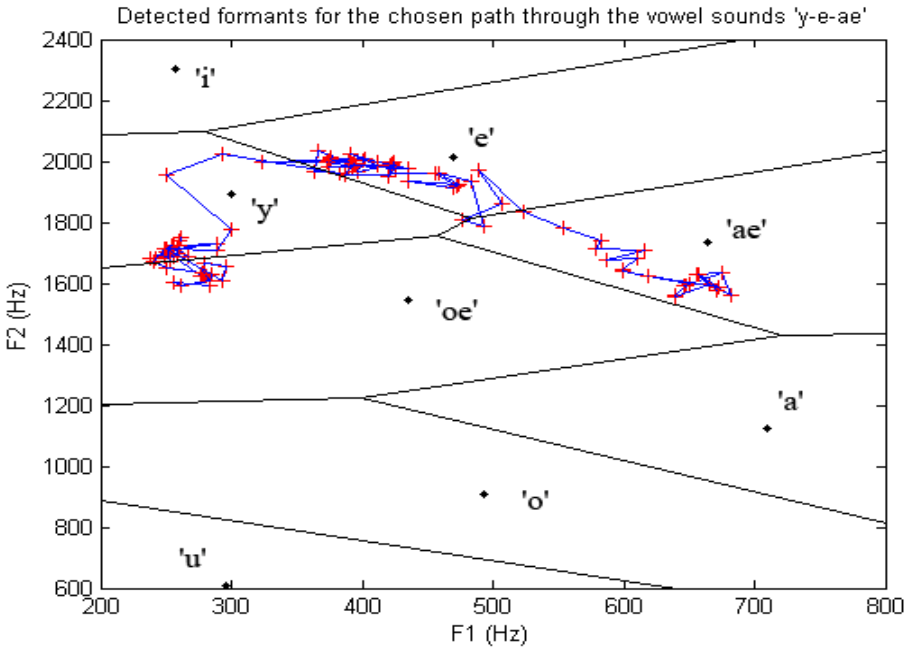


Figure 32. Formant frequencies corresponding to the final detected vocal tract shapes of the transition ”/y-e-ae/” (plus signs) with the trajectory through the formant space (solid line).

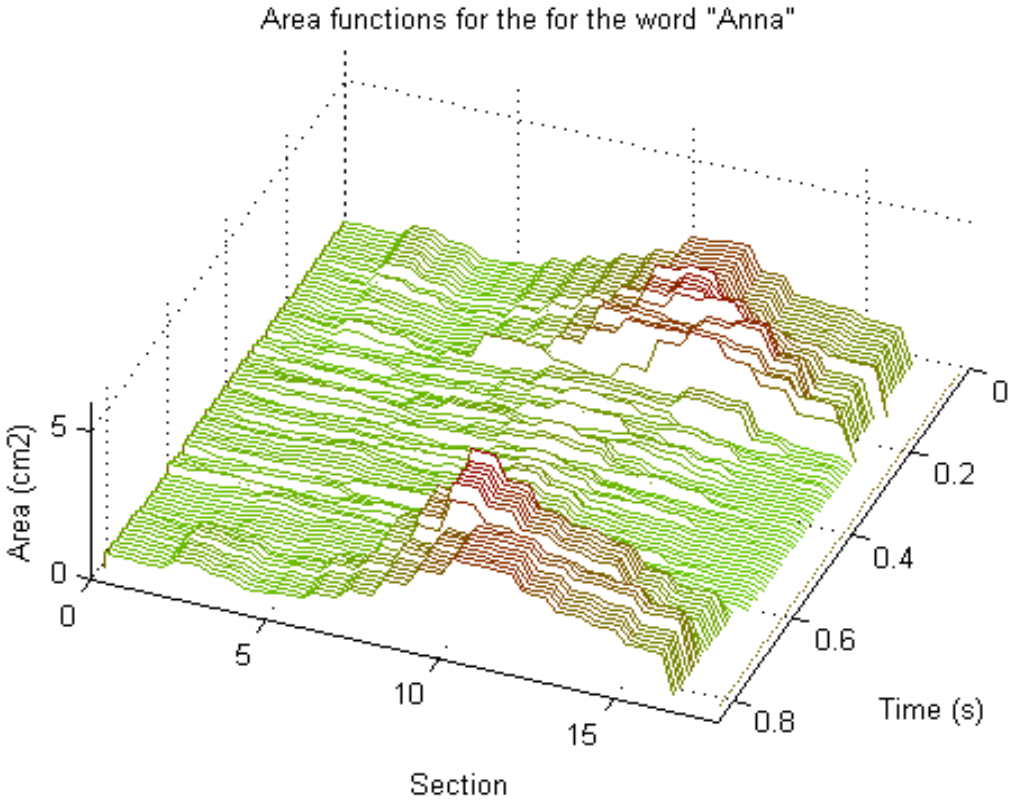


Figure 33. Area functions for the word “Anna”.

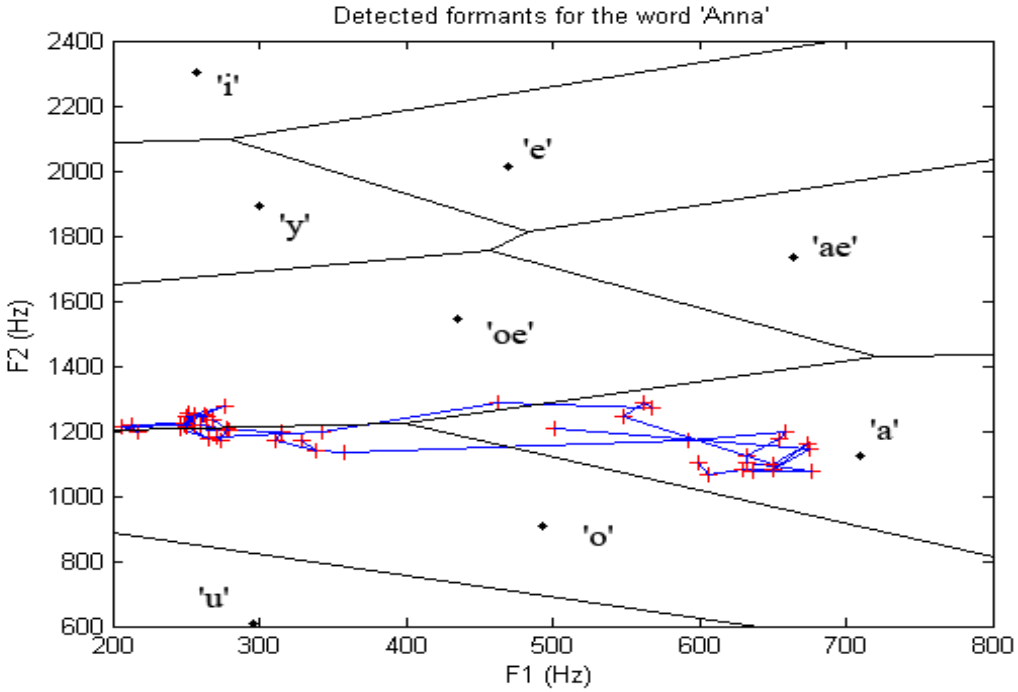


Figure 34. Formants and their connectors for the word "Anna".

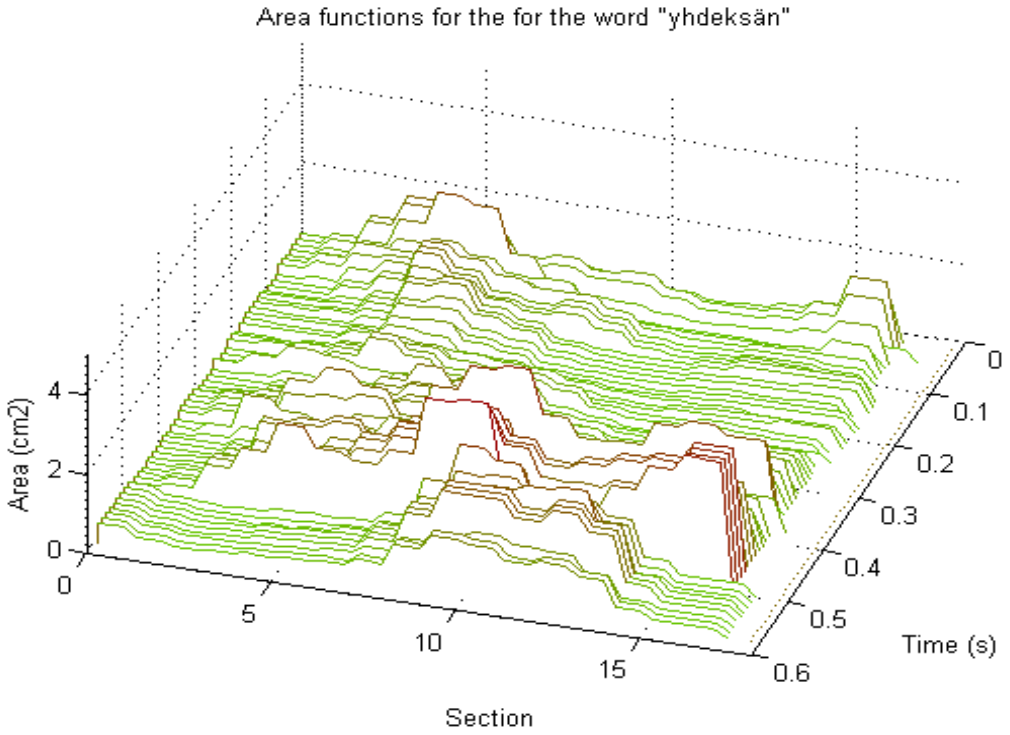


Figure 35. Area functions for the word “yhdeksän”.

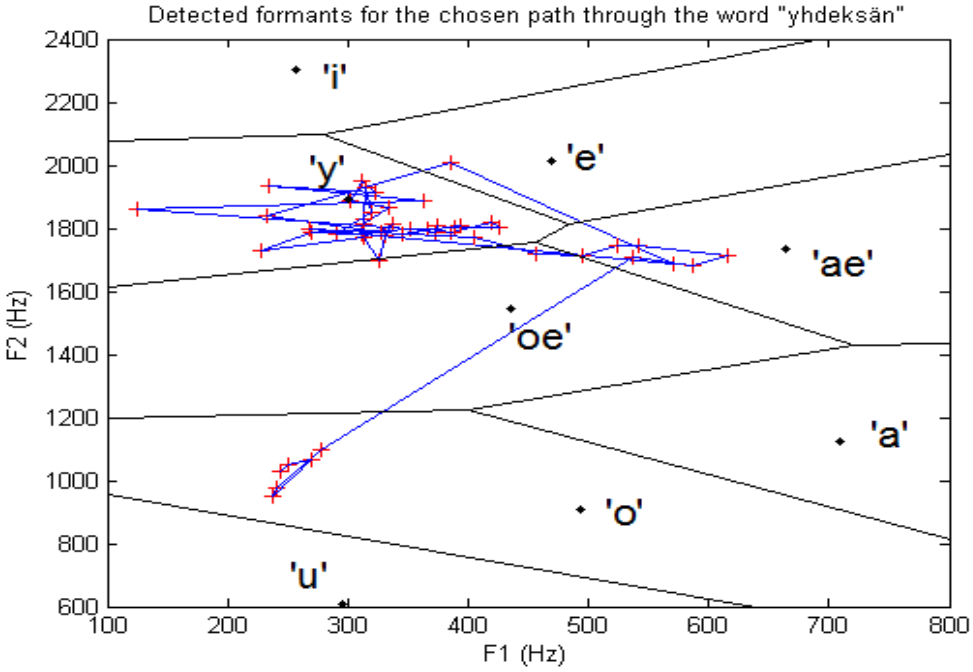


Figure 36. Detected formants and their connectors for the word "yhdeksän".

In the next test, a glide consonant /j/ was tested using a word “joi” ([joi]). The result can be seen in Figure 37 and Figure 38. /j/ is very close to the vowel /i/ but has a very high constriction in the middle of the mouth. In the formant plot, both sounds are detected to the same formant region around the vowel sound /i/. However, the difference between /j/ and /i/ sounds can be seen in the waterfall plot. This demonstrates nicely the potential of the shape trajectories in differentiation of speech sounds that are difficult to tear apart using classical instantaneous spectral representations. In /j/ the constriction starts a little bit deeper in the oral cavity than in /i/. The transition to /o/ can be seen as the extended part in the back of mouth gliding towards the front.

The final test is performed with the fricative consonant /v/ in the word “voi” ([v][o][i]; Figure 39) It can be observed that during the first frames the entire vocal tract has very low cross-sectional areas. This is due to the fact that the same formant frequencies are again obtained with the tract being scaled with different constants. In this particular case, a scaled-down shape is chosen by the optimization algorithm. Transition to /o/ needs the expansion of the front part of the vocal tract, and in /v/ the front part of the tract is constricted. Less energy is required, when the transition to /o/ is made from an area function with overall low values, than when the back of the vocal tract would have to be constricted as well.

In order to further illustrate this effect, the same shapes are shown in Figure 40 but for visualization every area function is scaled so that its maximum value is strictly at 4 cm^2. One can see that the tract is more expanded at the back and heavy constriction takes place close to the lips. If accurate information regarding the vocal tract shape itself is desired, and less attention is paid to the absolute values of maximum and minimum constriction, this type of strict scaling may be used to get the overall vocal tract shape during some consonant sounds as well. In Figure 41 , the formants of the word “voi” are shown. For the fricative /v/, the formant frequencies are found similarly to /n/ from the region of (200 Hz, 1200 Hz) for the first two formants.

In general, the experiments show that the tracking method provides valuable information of vocal tract movements during voiced speech. The obtained trajectories are somewhat disturbed during skipped time frames in some consonants during tracking of complete words, but at least glides, fricatives, and nasals can be tracked to a degree. Scaling the detected shapes to a fixed area range may be a good idea when the overall shapes are of interest. Better smoothing algorithms for the continuous tract shape would also enhance visualization of the data.

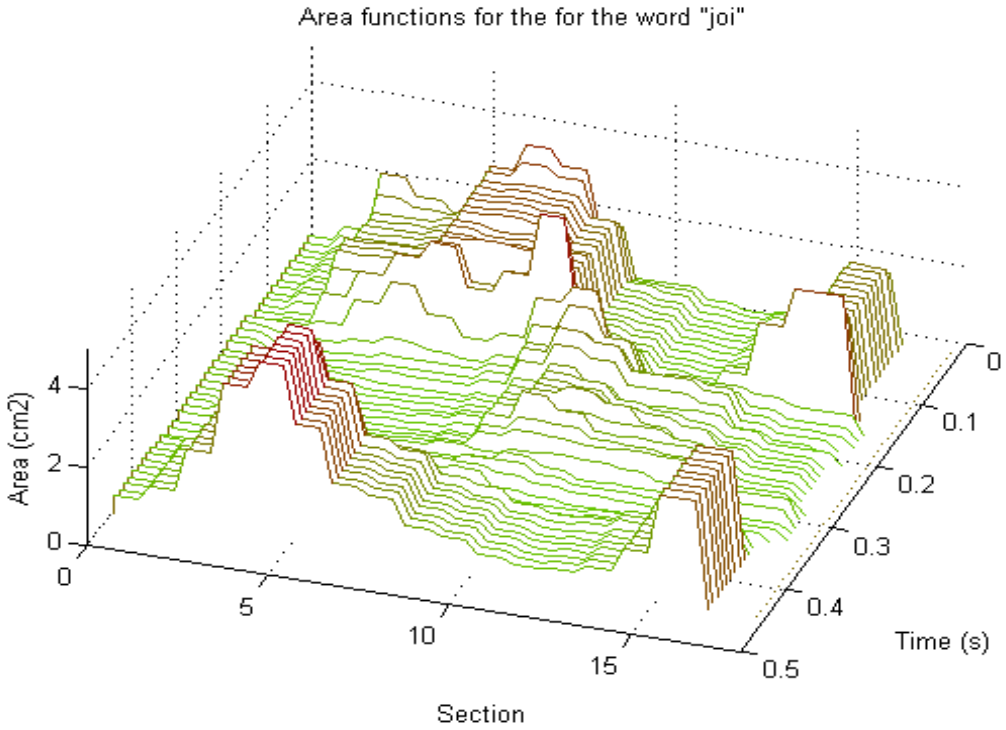


Figure 37. Area functions for the word “joi”.

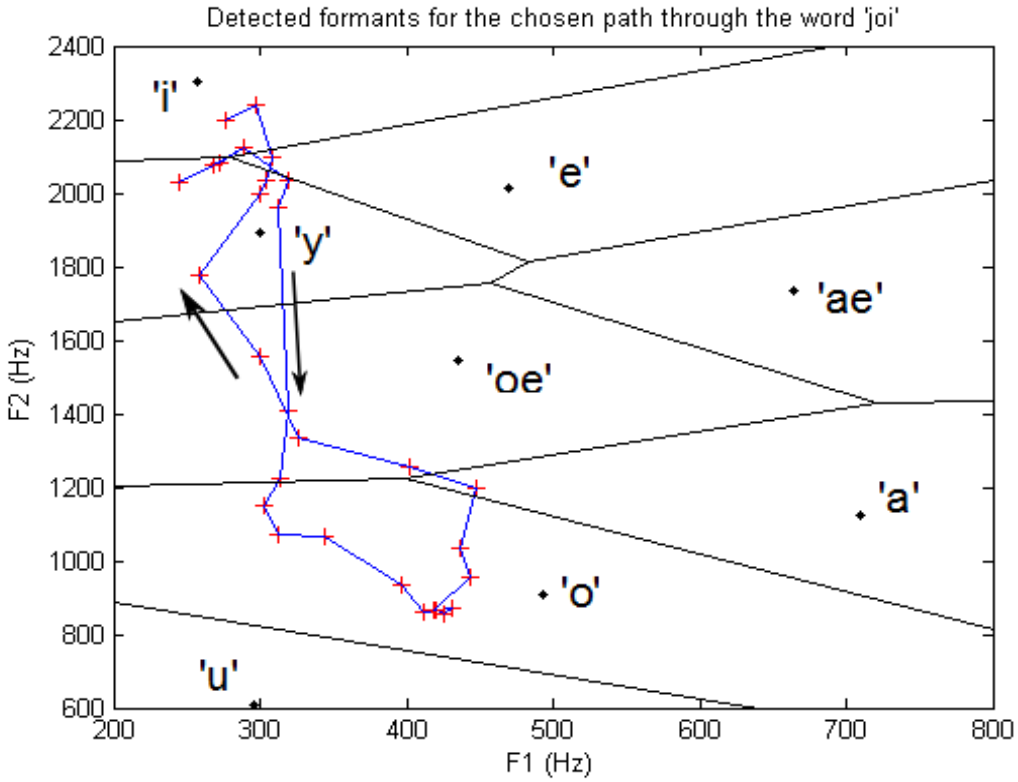


Figure 38. Detected formants and their connectors for the word "joi"

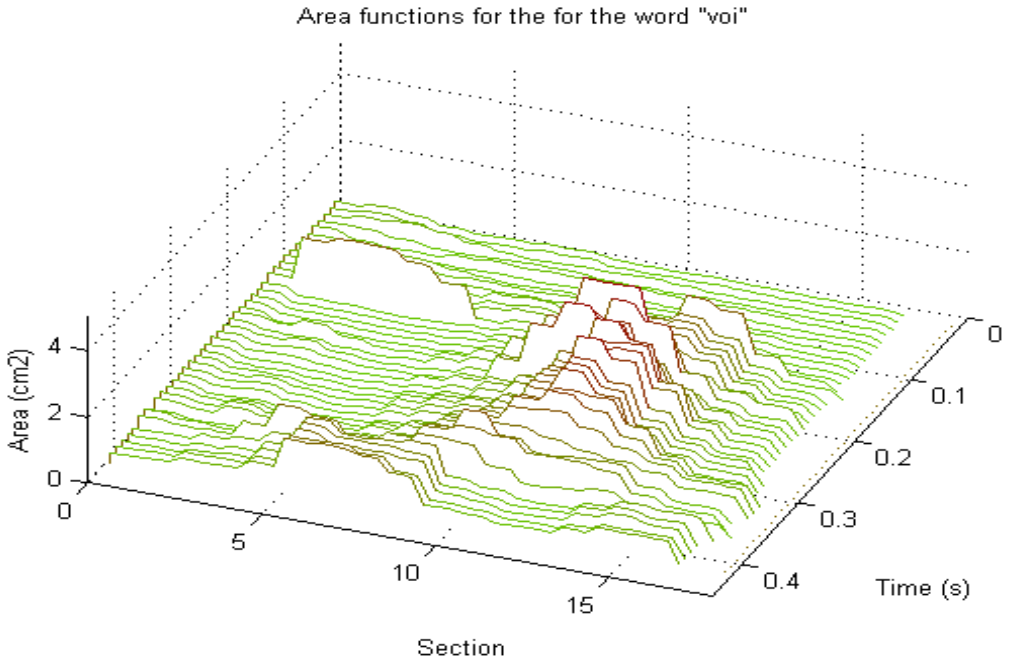


Figure 39. Area functions for the word "voi" as given from the tracker without scaling

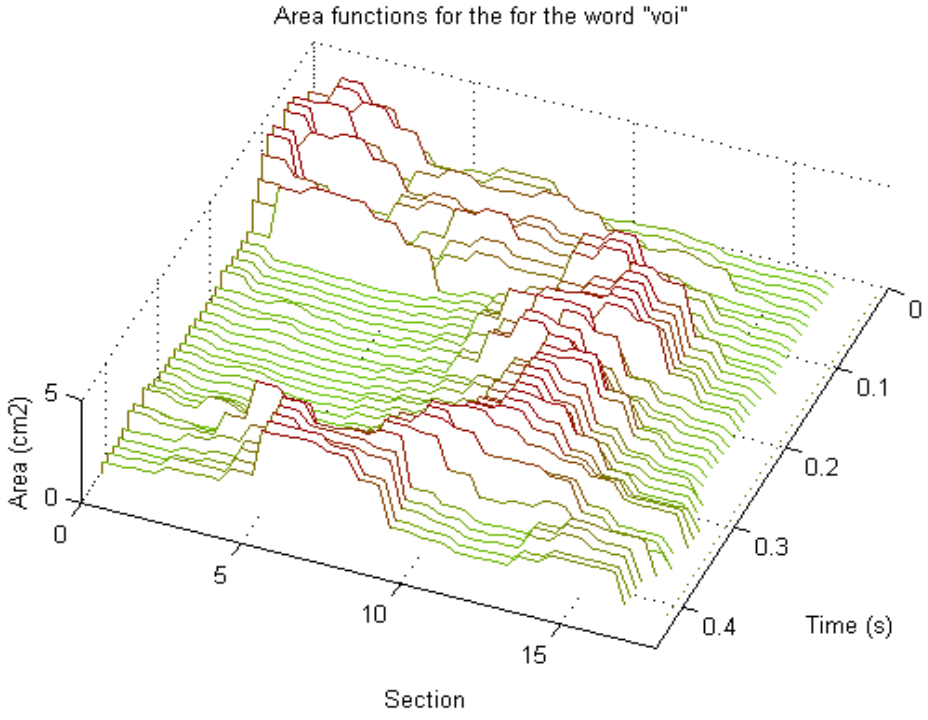


Figure 40. Area functions for the word "voi". The same shapes are used as in the previous figure but at every time window the vocal tract is scaled by a constant so that the maximum value is at 4 cm^2

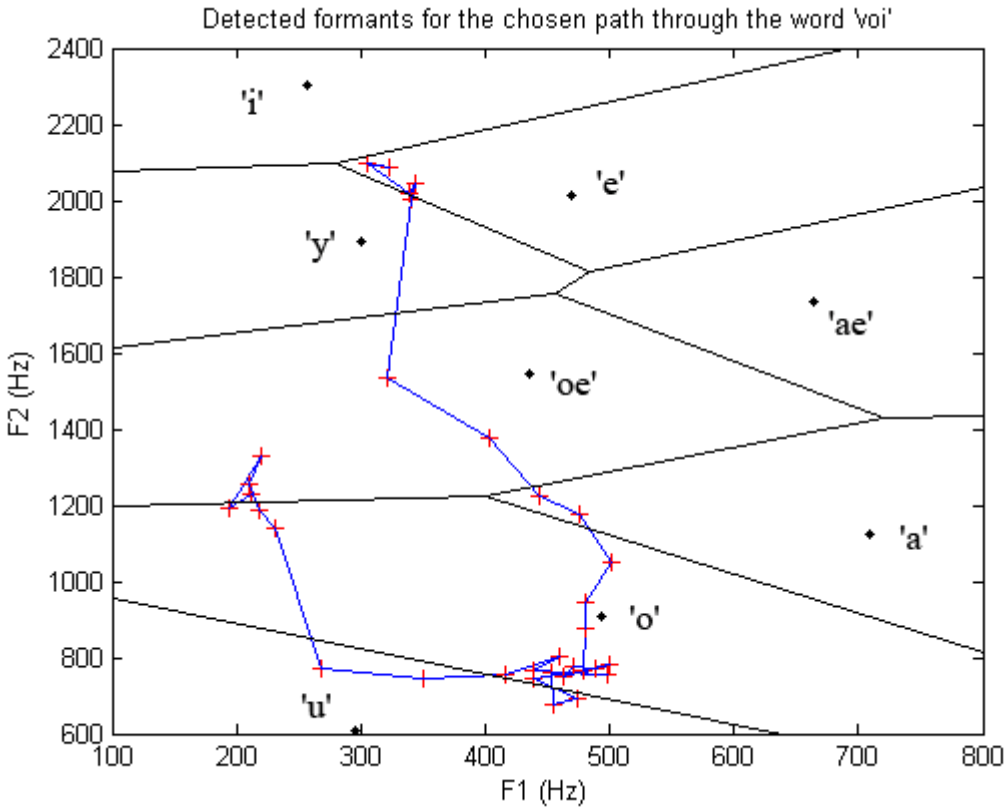


Figure 41. Detected formants and their connectors for the word "voi".

7 Conclusions and challenges for future research

The vocal tract shapes and shape trajectories obtained with the proposed method are yet to be verified by actual measurements. Based on the preliminary experiments with vowel sounds, it can already be said that the method gives promising results at least when dealing with voiced oral sounds. The area function tracking gives realistic trajectories for vowel chains and for some types of consonants as well. The trajectory optimization algorithm reduces heavily the amount of unrealistic vocal tract shapes of continuous speech signals. As soon as the physiological imaging techniques become sufficiently advanced for obtaining vocal tract shapes and speech signals simultaneously, the physical validity of the automatically detected vocal tract shapes can be verified.

Throughout the development of this work, many new questions emerged and many problems and results require further investigation. The models and methods used could be also improved further. One of the main principles in this work was to keep the number of parameters as low as possible. In the end, the vocal tract model was implemented using only 17 parameters. When changes were created to the area functions during the mapping phase, combinations of only five variables were required to obtain all of the vocal tract shapes.

During examination of the resulting vocal tract shapes, it was noticed that the scaling property of the vocal tract area functions (see section 5.3.2) may occasionally cause problems. A characteristic property of the vocal tract area function is that when the entire area function is scaled up or down by a constant factor, the formant frequencies stay relatively unchanged. Only a little change on the spectrum is created by the lip radiation impedance, which depends on the lip area that is scaled simultaneously. Similar formant frequencies for shapes with different scales lead to the effect in which the mapping includes series of shapes that are scaled versions of each other.

Due to this scaling property, the area function tracking can as well lock into one of the scaled versions of the tract instead of the desired real physiologically motivated version. For instance, when tracking vowel sound /a/, it is possible to find area functions with same formant characteristics, when the front part of the oral cavity reaches 12 cm^2 or 4 cm^2. The problem is partially solved by using a large amount of competing vocal tract shapes, when it becomes probable that the overall amount of change across all small area tract candidates is smaller than in the upscaled versions. This minimization is taken care by the shortest path algorithm. On the other hand, this may lead to detection of too small areas, as occurred in the experiments in the case of the fricative /v/. The profile scaling affects the lip opening area and the formant bandwidths. It may be possible to add the bandwidth information to the estimation process and to search for the optimal scales as well.

Also another idea for solving the problem of heavily scaled shapes occurred during the analysis of the results. During the variation phase, all created vocal tract areas could be scaled to a desired area range with a fixed maximum opening. In addition to removing the scaling problem, this could also reduce the number of points required in the lookup table since redundancy would be limited. The overlapping points in the mapping could thus be deleted. However, this method would cause further problems in the actual shape trajectory estimation. It is physiologically possible to produce some

sounds with tract shape scaled in different manners, and this property is most likely used in transitions between sounds.

An interesting quality was noticed in the created mapping as well. The straight small-density line seen in the density figure tells that the vowel sounds do not preferably enter this region. Namely, it divides the vowel sounds in two categories of back and front vowels. Since back and front vowels are physiologically classified according to their point of highest constriction in the vocal tract, the position of the small density line changes as a function of the vocal tract length. This raises a question whether it would be possible to estimate the vocal tract length of a speaker using the formant density information, which would be useful in many speech processing applications dealing, e.g., with speaker normalization. However, it is possible that this kind of a histogram on the F1-F2 space is an artifact coming from the KL-variant and the method to create the variability around the anchor points. More research is needed to solve this question.

Some improvements to the acoustic-to-articulatory mapping could also be made. The sensitivity functions could be calculated again at every iteration using the technique described by Fant and Pauli [32]. This would lead to more orthogonal movements of the formant frequencies. Also, different ways of locking the glottis in place could be implemented to get more versatile variations to the area functions. Some tests were made by simply changing the area of the first tube section at the glottis after shape modulation to its original value and leaving the DC-term untouched, but this actually made the formant movements less orthogonal.

Regarding the Kelly-Lochbaum model itself, the viscous and thermal losses could be taken into account as well, possibly leading to slightly more realistic vocal tract transfer functions, although the lip radiation impedance considered in this work is the biggest source of losses of the vocal tract, except the low F1 which attenuates much due to the yielding walls [39]. The vocal tract model was designed so that the length of the tube at the glottal end could be also varied, but for this work the feature was not used for the sake of parameter reduction. In more sophisticated models the length of every individual uniform section may also be changed using fractional delays, but in order to keep the simulations as simple as possible, this work did not comprise such a property. In the future versions also the nasal tract should be considered. Also, we should be able to model cases, where the sound source is not at the glottis, e.g., fricatives (like /s/) and plosives.

A lot of research on tracking the articulatory gestures from speech signals is being carried out today. Advanced vocal tract models and powerful computers for modeling give researchers extremely valuable tools that could only be dreamed of at the early stages of speech research. Similar research methods are used to fill knowledge gaps in different fields of research, such as understanding the evolution of human vocal apparatus for communication purposes, and the process of children language acquisition and communication skills of a child. This thesis gave merely a scratch on the wide research field, but introduced, and participated in solving, some of the main problems faced by researchers today in understanding human speech production, the most delicate of all forms of communication.

References

[1] S. Tamura, K. Iwano, and F. Sadaoki, "Towards robust multimodal speech recognition," Department of Computer Science, Tokyo Institute of Technology, 2005.

[2] M. Studdert-Kennedy, "On learning to speak," *Human Neurobiology*, no. 2, pp. 191-195, 1983.

[3] M. Dolar, *A voice and nothing more*. Cambridge, MA: MIT Press, 2006.

[4] H. Dudley, R. R. Riesz, and S. A. Watkins, "A Synthetic Speaker," *J. Franklin Inst. 227*, pp. 739-763, 1939.

[5] J. L. Flanagan, "Voices of Men and Machines," *J. Acoust. Soc. Am.*, vol. 51, no. 5, pp. 1375-1387, Mar. 1972.

[6] H. Dudley, "The Vocoder," *Bell Labs Rec.*, vol. 18, pp. 122-126, Dec. 1939.

[7] J. L. Flanagan, K. Ishizaka, and K. L. Shipley, "Signal models for low bit-rate coding of speech," in *J. Acoust. Soc. Am.*, 1980, pp. 780-791.

[8] J. Schroeter and M. M. Sondhi, "Techniques for Estimating Vocal-Tract Shapes from the Speech Signal," *IEEE Trans. Speech, Audio Processing*, vol. 2, no. 1, pp. 133-150, Jan. 1994.

[9] J. Shroeter, J. N. Larar, and M. M. Sondhi, "Speech Parameter Estimation Using a Vocal Tract/Cord Model," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 308-311, Apr. 1987.

[10] P. Mermelstein, "Articulatory model for the study of speech production," *J. Acoust. Soc. Am.*, vol. 53, no. 4, pp. 1070-1082, 1973.

[11] J. N. Larar, J. Schroeter, and M. M. Sondhi, "Vector quantization of the articulatory space," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 12, pp. 1812-1818, 1988.

[12] B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey, "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer sorting technique," *J. Acoust. Soc. Am.*, vol. 63, no. 5, pp. 1535-1555, 1978.

[13] M. Stone, "Imaging the tongue and vocal tract," *International Journal of Language & Communication Disorders*, vol. 26, no. 1, pp. 11-23, 1991.

[14] J. Malinen and P. Palo, "Recording speech during MRI: Part II," *Proceedings of the 6th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2009.

[15] Y. Laprie and B. Mathieu, "A variational approach for estimating vocal tract shapes from the speech signal," *Proceedings of the ICSLP'98*, vol. 2, pp. 929-932, May 1998.

[16] S. Maeda, "Compensatory articulation during speech: evidence from the analysis of vocal tract shapes using an articulatory model"," in *Speech production and speech modeling*. Kluwer Academic Publishers, 1990, pp. 131-149.

[17] J. Dang and K. Honda, "Estimation of vocal tract shapes from speech sounds with a physiological articulatory model," *Journal of Phonetics*, vol. 30, pp. 511-532, 2002.

[18] R. Carré, "From acoustic tube to speech production," *Speech Communication 42*, pp. 227-240, 2004.

[19] P. Badin and G. Fant, "Notes on the vocal tract computations," *STL-QPSR*, vol. 25, no. 2-3, pp. 53-108, 1984.

[20] R. Carré, "Dynamic properties of an acoustic tube: Prediction of vowel systems," *Speech Communication*, vol. 51, no. 1, pp. 26-41, Jan. 2009.

[21] J. L. Flanagan, *Speech Analysis Synthesis and Perception*, 2nd ed. Berlin: Springer-Verlag, 1972.

[22] K. Wiik, *Fonetiikan perusteet*, 2nd ed. Porvoo / Helsinki / Juva: WSOY, 1981.

[23] G. Fant, *Acoustic theory of speech production*, 2nd ed. The Hague, The Netherlands: Mouton & Co., 1970.

[24] M. M. Sondhi, "Resonances of a bent vocal tract," *J. Acoust. Soc. Am*, vol. 79, no. 4, pp. 1113-1116, Apr. 1986.

[25] H. W. Strube, "The meaning of the Kelly-Lochbaum acoustic-tube model," *J. Acoust. Soc. Am.*, vol. 108, no. 4, pp. 1850-1855, Oct. 2000.

[26] J. L. Kelly and C. C. Lochbaum, "Speech Synthesis," *Proc. 4th Int. Congr. Acoustics, Copenhagen*, pp. 1-4, Sep. 1962.

[27] T. I. Laakso, V. Välimäki, M. Karjalainen, and U. K. Laine, "Splitting the Unit Delay, Tools for fractional delay filter design," *IEEE Signal Processing Magazine*, vol. 13, no. 1, pp. 30-60, Jan. 1996.

[28] U. K. Laine, "Modelling of Lip Radiation Impedance in z-domain," *IEEE Int. Con. Acoust., Speech, Signal Processing*, vol. 3, May 1982.

[29] P. M. Morse and K. U. Ingard, *Theoretical acoustics*. New York: McGraw-Hill, 1968.

[30] P. Haughton, *Acoustics for audiologics*, 1st ed. the USA: Academic Press, 2002.

[31] D. Hill, L. Manzara, and C. Schock, "Real-time articulatory speech-synthesis-by-rules," *Proceedings of AVIOS '95, San Jose*, pp. 27-44, Nov. 1995.

[32] G. Fant and S. Pauli, "Spatial characteristics of vocal tract resonance modes," *Proc. Speech Communication Seminar, Stockholm*, pp. 121-132, 1974.

[33] L. R. Rabiner and R. W. Schafer, "Introduction to Digital Speech Processing," *Foundations and Trends(R) in Signal Processing*, vol. 1, no. 1-2, pp. 1-194, 2007.

[34] B. H. Story, I. R. Titze, and E. A. Hoffman, "Vocal tract area functions from magnetic resonance imaging," *J. Acoust. Soc. Am*, vol. 100, no. 1, pp. 537-554, Jul. 1996.

[35] K. Wiik, "Finnish and English Vowels," University of Turku, Doctoral dissertation, 1965.

[36] C. Ericsdotter, "Articulatory-Acoustic Relationships in Swedish Vowel Sounds," Stockholm University, Department of Linguistics, Doctoral Dissertation, ISBN: 91-7155-151-4, 2005.

[37] J. R. Deller, J. G. Proakis, and J. H. Hansen, *Discrete-Time Processing of Speech Signals*. New York: Macmillan Publishin Company, 1993.

[38] R. E. Bellman and S. E. Dreyfus, "Applied Dynamic Programming," in *Princeton University Press*, Princeton, NJ, 1962.

[39] J. Harrington and S. Cassidy, *Techniques in Speech Acoustics*. The Netherlands: Kluwer Academic Publishers, 1999.

[40] P. Palo, "A Review of Articulatory Speech Synthesis," Master's thesis, Helsinki University of Technology, Department of Electrical and Communications Engineering 2006.

Appendix A

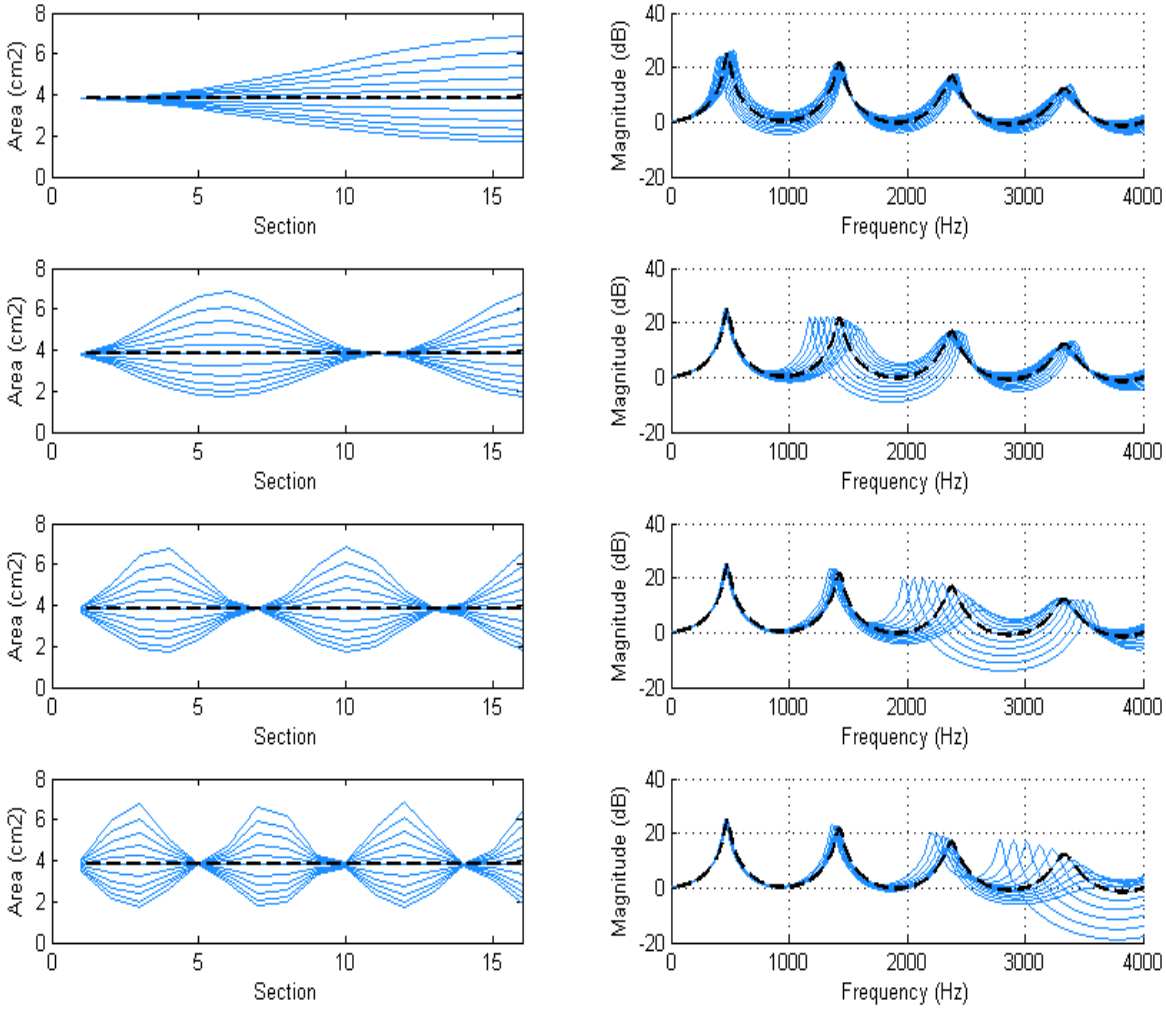


Figure A-1. Variations created to a uniform tube using 5 iterations towards more constricted and 5 iterations towards less constricted shapes for each formant. The variation is done by creating a generating function by adjusting the weighting factor of a term of discrete cosine series, and using this in modulating the profile stepwise. The four first even coefficients are adjusted separately, and the result is shown in the figures on the left. Corresponding magnitude spectra are shown in the figures on the right. Light line shows all the profiles and their spectra created. The dark dashed line shows the original shape and its spectrum. It is noticed that in the case of the uniform tube the variations in each cosine term cause remarkably orthogonal changes in the corresponding formant frequency.

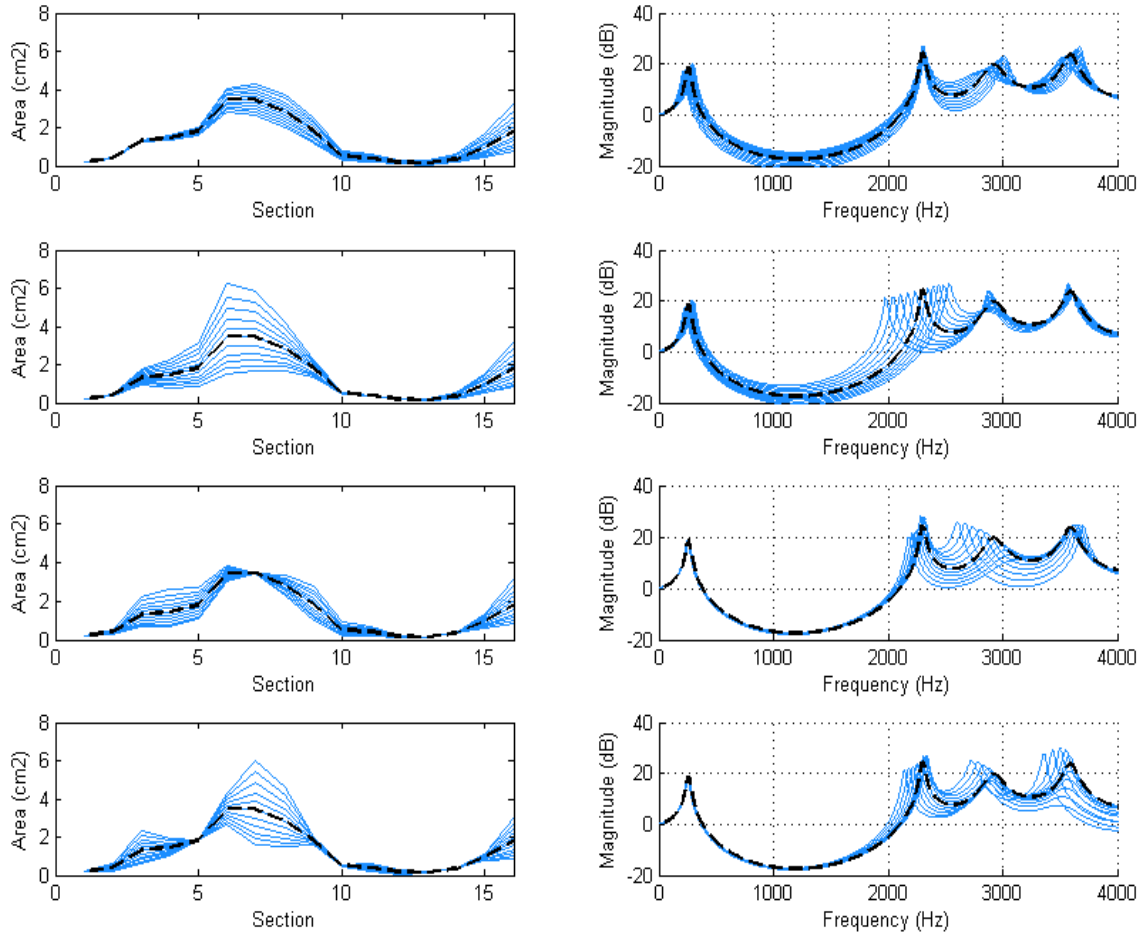


Figure A-2. Variations created to the area function of the Finnish vowel sound /i/ using same method as in Figure A-1. The first formant stays almost in its original location. The degree of constriction in the mouth cavity is high, and only a small amount of variation is obtained with the modulation. The second formant again is allowed to shift more freely, because the tract is more open in the middle of the tract and at the front part of the mouth.

Appendix B

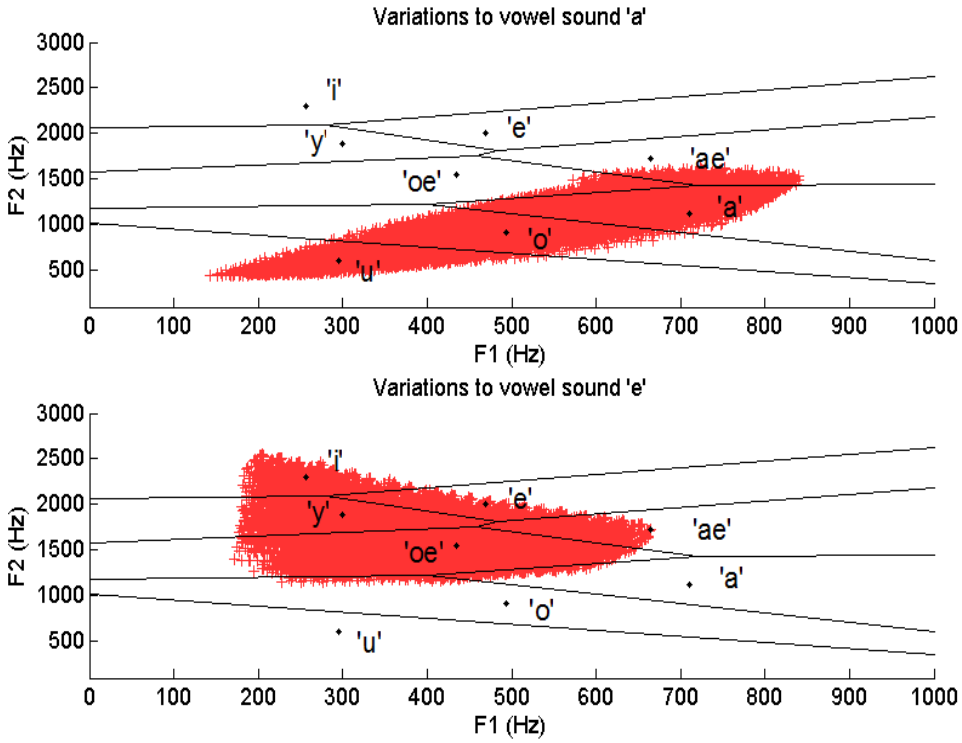


Figure B-1. Formants 1 and 2 for variations of vowel sounds /a/ and /e/.

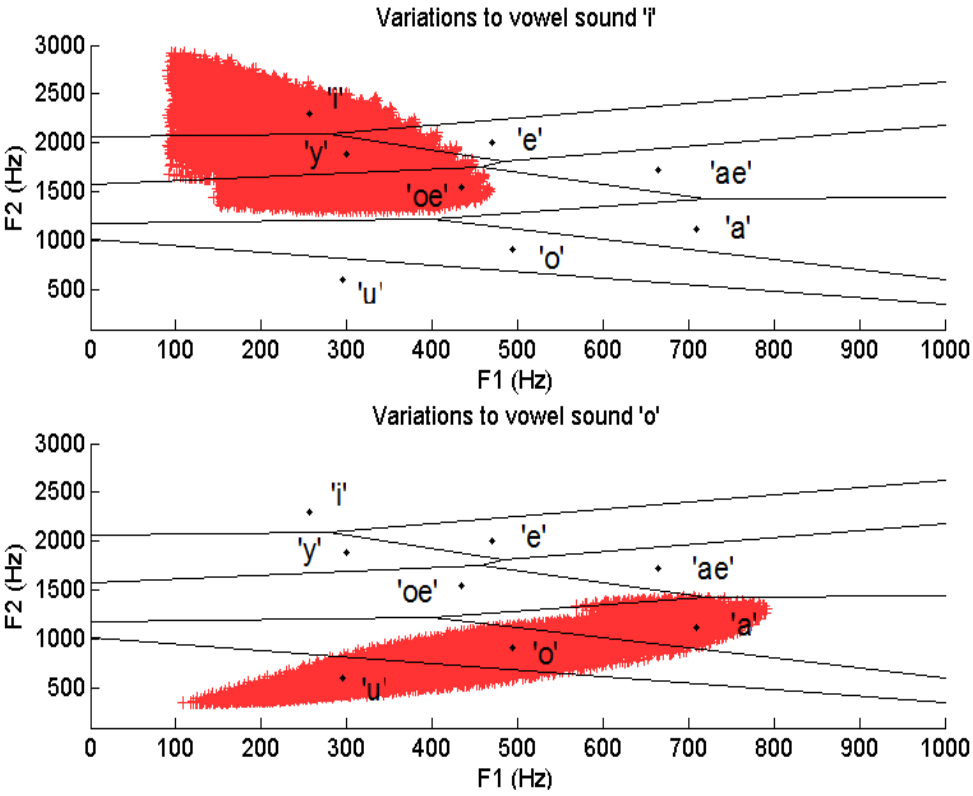


Figure B-2. Formants 1 and 2 for variations of vowel sounds /i/ and /o/.

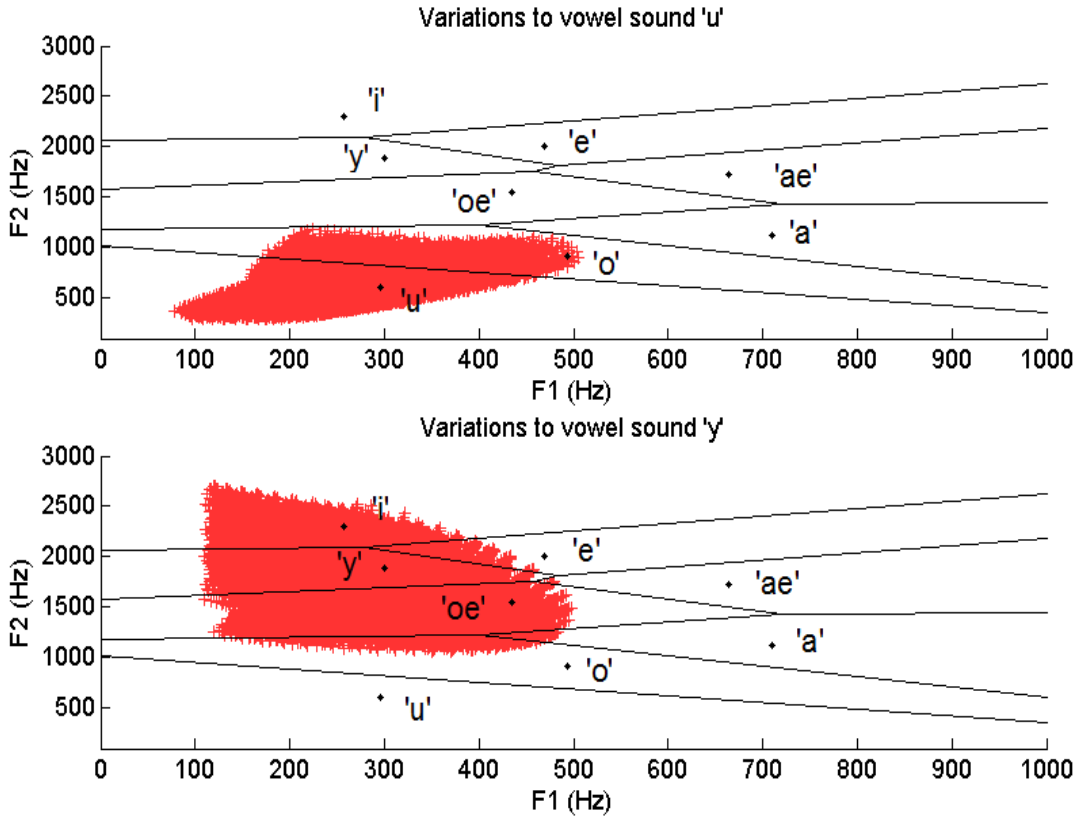


Figure B-3. Formants 1 and 2 for variations of vowel sounds /u/ and /y/.

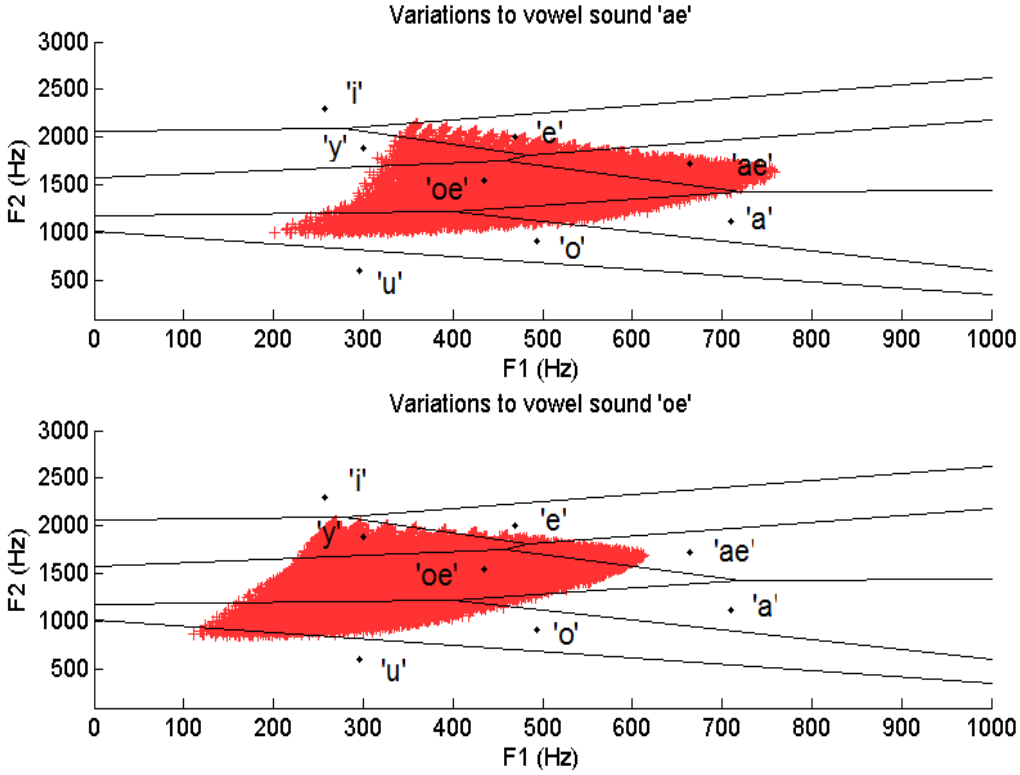


Figure B-4. Formants 1 and 2 for variations of vowel sounds /ae/ and /oe/.