USC–SIPI REPORT #439

VISUALIZING AND MODELING VOCAL PRODUCTION DYNAMICS

by

Erik Bresch

May 20ll

Signal and Image Processing Institute
UNIVERSITY OF SOUTHERN CALIFORNIA
USC Viterbi School of Engineering
Department of Electrical Engineering-Systems
3740 McClintock Avenue, Suite 400
Los Angeles, CA 90089-2564 U.S.A.

VISUALIZING AND MODELING VOCAL PRODUCTION DYNAMICS

by

Erik Bresch

A Dissertation Presented to the
FACULTY OF THE USC GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(ELECTRICAL ENGINEERING)

May 2011

Table of Contents

List Of Tables

List Of Figures

Abstract

Understanding human speech production is of fundamental importance for basic and applied research in human communication: from speech science and linguistics to clinical and engineering development. While the vocal tract posture and movement can be investigated using a host of techniques, the newly developed real time (RT)-magnetic resonance imaging (MRI) technology has a particular advantage - it produces complete views of the entire moving vocal tract including the pharyngeal structures in a non-invasive manner. RT-MRI promises a new means for visualizing and quantifying the spatio-temporal articulatory details of speech production and it also allows for exploring novel data-intensive, machine learning based computational approaches to speech production modeling.

The central goal of this thesis is to develop new technological capabilities and to use these novel tools for studying human vocal tract shaping during speech production. The research, which is inherently interdisciplinary, combines *technological* elements (to design engineering methods and systems to acquire and process novel speech production data), *experimental* elements (to design linguistically meaningful studies to gather useful insights) and *computational* elements (to explain the observed data and design predictive capabilities).

In Chapter 1, which was in part published in [6], the use of RT-MRI as an emerging technique for speech production research studies is motivated. An outline is provided

of the biomedical image acquisition and image processing challenges, potentials, and opportunities arising with the use of RT-MRI.

The second part, Chapter 2, describes novel hardware technology and signal processing algorithms which were developed to facilitate synchronous speech audio recordings during RT-MRI scans. Here, the main problem lies in the loud noise produced by the MRI acquisition process. The proposed solution incorporates digital synchronization hardware and an adaptive signal processing algorithm which allows the acquisition of speech audio with satisfactory quality for further analysis. This enables joint speech-image data acquisition that in turn allows for joint modeling of articulatory-acoustic phenomena. Most of this chapter was published in [9].

Subsequently, Chapter 3 addresses the extraction of relevant geometrical features from the vast stream of magnetic resonance (MR) images. In the case of the commonly used midsagittal view of the human vocal tract the geometrical features of interest are the locations of the articulators, and hence the underlying image processing problem to be solved is that of edge detection. Further complications arise from the poor MR image quality, which is compromised by the inherent trade-off between spatial, temporal resolution, and signal to noise ratio. A solution to the edge detection problem will be devised using a deformable geometrical model of the human vocal tract. Mathematically the proposed procedure relies on designing alternate gradient vector flows for the solution of a non-linear least squares optimization problem. With the new method the human vocal tract outline can be traced automatically. These findings were published in [7].

Chapters 4 and 5 describe two vocal production studies using articulatory vocal tract data. The first study investigates 5 soprano singers' static vocal tract shaping during the singing production of vowel sounds, and it considers the much-researched theory of resonance tuning. The study successfully validates the usefulness of RT-MRI data and the

data processing methods of Chapters 2 and 3. The second study focuses on the tongue shaping of English sibilant fricative sounds, and reproduces previously known findings with the new RT-MRI modality. The findings of these two studies have been published in [8, 10].

The last part of this thesis is contained in Chapter 6 and it proposes a statistical framework for the modeling of articulatory speech data. Here, the main focus lies on the coupled hidden Markov model (CHMM) as a candidate system to capture the dynamics of the multi-dimensional vocal tract shaping process. It is demonstrated that using this methodology it is possible to capture in a data driven way the well-known timing signatures of the velum-oral coordination of English nasal sounds in word onset and coda positions. The content of this chapter has been published in [5].

This thesis is concluded with a brief summary of the contributions and a discussion of possible future research directions in Chapter 7.

Chapter 1

Introduction to speech production research using RT-MRI

Understanding human speech production is of great interest from engineering, linguistic, and several other research points of view. While several types of data that are available to speech production studies lead to different avenues for research, in this section we focus on RT-MRI, as an emerging technique. We discuss the details and challenges of RT-MR acquisition and analysis, and modeling approaches that make use of MRI data for speech research.

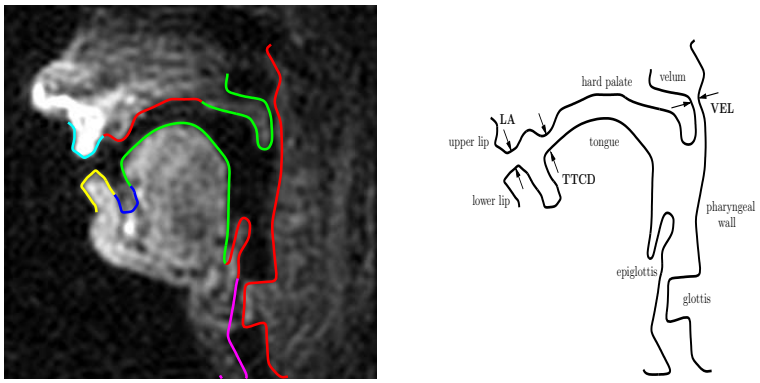
From an engineer's point of view, detailed knowledge about speech production gives rise to refined models for the speech signal that can be exploited for the design of powerful speech recognition, coding, and synthesis systems. From the linguist's point of view, speech research may be conducted to address open questions in the areas of phonetics and phonology. These include, what articulatory mechanisms explain the inter and intra subject variability of speech, what aspects of the vocal tract shaping are critically controlled by the brain for conveying meaning and emotions, and how does prosody affect the articulatory timing. From other research points of view, speech production is important to understand language acquisition and language disorders. All of these efforts require intimate knowledge of the speech generation mechanisms.

Different types of data are available to the speech researcher – from audio and video recordings of speech production to muscle activity data produced by electromyography, respiratory data from subglottal or interoral pressure transduction, and images of the larynx obtained through video laryngoscopy. Table 1.1 summarizes the most commonly used techniques and lists advantages and disadvantages. While the vocal tract posture and movement can be investigated using a number of techniques including x-ray (microbeam), cinefluography, ultrasound, palatography, electromagnetometry (EMA), RT-MRI has a particular advantage – it produces complete views of the entire vocal tract including the pharyngeal structures. Furthermore, RT-MR is a non-invasive and safe procedure. It allows to capture 3-dimensional data, and also flesh-point tracking is possible using a variant of this technology called tagged-MRI.

However, RT-MRI also poses a number of technical challenges. The most important one for the speech researcher lies in the relatively poor spatial and temporal resolution of the method. While the MR technology inherently allows a trade-off between spatial and temporal resolution the generally desired sub-mm accuracy at frame rates of 200 frames per second far exceeds what is technically possible today [45]. The MRI technology is also relatively expensive when compared with other data acquisition methods, and it is limited to subjects without major dental work or implants. Moreover a supine position is generally required for the subject, which may or may not affect the vocal tract shaping processes. Lastly, simultaneous audio recordings are difficult to obtain, mainly due to the loud scanner noise.

With RT-MRI, a midsaggital image of the vocal tract from the glottis (bottom) to the lips (left) can be acquired as illustrated in Figure 1.1(a). In this image, we can trace the air-tissue boundaries of the anatomical components that are of interest to the speech researcher and obtain a representation similar to Figure 1.1(b). These components, also

known as articulators, are controlled by the brain during speech production and are used to change the shape of the vocal tract tube. With it, they also change the filter function for the excitation signal generated at the glottis and elsewhere along the airway. Hence the motion of the articulators shapes the sounds of speech and other human vocalizations.



(a) Midsagittal real-time MR image with contours of interest.

(b) Articulators and sample vocal tract variables.

Figure 1.1: Example vocal tract MR image and tract variables

The signal processing challenges when studying speech production using RT-MRI lie in the fast acquisition of high-quality RT-MR images including simultaneous noise-robust audio recording, the subsequent detection of the relevant features from each image, and the analysis and modeling of the time-varying vocal tract shape for the purpose of gaining deeper understanding of the underlying principles that govern the speech production process.

Table 1.1: Methods for acquiring speech production data of the vocal tract

Method	Pros	Cons	Comments
CT	• high temporal and spatial resolution • captures pharyngeal structures • 3–D possible	• exposure to radiation	• rarely used in speech research
EMA	• high spatial and temporal resolution • 3–D	• provides spatially sparse point tracking data • cannot capture pharyngeal structures	• often used in speech production studies.
x-ray (microbeam)	• high spatial and temporal resolution • flesh point tracking not possible for pharyngeal structures	• exposure to radiation • images show only a projection through volume which makes contour extraction difficult • x-ray microbeam equipment not widely available • spatially sparse data	• rarely used now in speech research • existing data bases are still being used
Ultrasound	• high temporal resolution • non-invasive, safe • good audio can be obtained simultaneously	• noisy images • detects only first tissue-air boundary • not suitable for anterior tongue tip and lip imaging • detector is in contact with jaw and may affect speech production process	• used primarily for tongue body imaging
MRI	• non-invasive, safe • captures pharyngeal structures • 3–D possible • tagged MRI allows flesh-point tracking	• relatively poor spatial and temporal resolution • expensive • limited to subjects without major dental work/implants • supine position generally required • simultaneous audio recording difficult due to scanner noise • teeth do not show in image	• an emerging technique for speech research

Chapter 2

Audio recordings during RT-MRI scans

2.1 Abstract

This chapter describes a data acquisition setup for recording, and processing, running speech from a person in an MRI scanner. The main focus is on ensuring synchronicity between image and audio acquisition, and in obtaining good signal to noise ratio to facilitate further speech analysis and modeling. An field programmable gate array (FPGA) based hardware design for synchronizing the scanner image acquisition to other external data such as audio is described. The audio setup itself features two fiber optical microphones and a noise-canceling filter. Two noise cancellation methods are described including a novel approach using a pulse sequence specific model of the gradient noise of the MRI scanner. The setup is useful for scientific speech production studies. Sample results of speech and singing data acquired and processed using the proposed method are given.

2.2 Introduction

In recent years, magnetic resonance imaging has become a viable tool for investigating speech production. Technological advances have enabled studying the structure of the

vocal tract, and its dynamical shaping, during speech production. For example, tongue deformation characteristics have been studied with a cine-MRI technique [57] and a real-time MR imaging technique described in [45] has been successfully used to capture the changing mid-sagittal shape of the vocal tract during speech production. One methodological challenge, however, is in synchronizing the acquisition of an audio signal with the collection of time-varying vocal tract images, which is important for any subsequent analysis and modeling of the acoustic-articulatory relation. In [57] the audio signal was recorded in a separate procedure after the MR images were collected so that synchronicity of the signals and images could be only approximately achieved through extensive training of the subject and with a restriction to few utterances.

There have been few studies where MR images and audio signals were obtained simultaneously. The problem is posed by the high intensity gradient noise caused by the scanner, which is in the audible frequency range. This degrades the audio signal such that acoustic analysis of the speech content is difficult, if not impossible. Previous studies such as [47] have addressed this problem using a correlation-subtraction method, where one captures the noise signal separately and relies on its stationarity. This method does not, however, account for non-stationary noise sources such as body movement of the subject or vibration of the cooling pump.

There are commercially available noise mitigation solutions that have been used in some MRI studies, such as the one by Phone-OR[1] [29] which provides an integrated MR-compatible fiber-optical microphone system that allows both real-time and offline noise cancellation. This proprietary system is described to use a special microphone assembly which houses two transducers, one to capture the speech signal and one to capture only the ambient noise. The two microphones are mounted in close proximity but their directional

[1] `http://phone-or.com`

characteristics are at a 90 degree angle so that one (main) microphone is oriented towards the mouth of the subject to capture the speech signal and the other (reference) microphone is oriented such that it rejects the speech signal and captures only the ambient noise. In our own experiments with this system, however, the reference signal contained a strong speech signal component and the subsequent noise cancellation procedure would remove the desired speech signal in addition to the noise to an extent that was undesirable for further analysis of the signal.

Hence, the initial research goal was the development of an alternative system in which a separate fiber optical microphone was located away from the subject and outside the magnet, but inside the scanner room, in a place where it captures almost exclusively the ambient noise and not the subject's speech. Importantly, also, this system was to capture the audio and the MR images simultaneously and to ensure absolute synchronicity for spontaneous speech and other vocal productions including singing. Subsequently, however, a better noise cancellation methodology was found which omits the use of a recorded noise reference signal altogether.

2.3 System level description of the data acquisition system

Figure 2.1 illustrates how the various components of the data acquisition system are located in the scan room, the systems room, and the control room of the MRI facility. Two fiber optical microphones are located in the scan room. The main microphone is approximately 0.5 inches (1.3 centimeters) away from the subject's mouth at a 20 degree angle, and the reference microphone is positioned on the outside of the magnet, roughly 3 feet (0.9 meters) away from the side wall at a height of about 4 feet (1.2 meters). The microphones connect to the optical receiver box, which is located in the MRI control room.

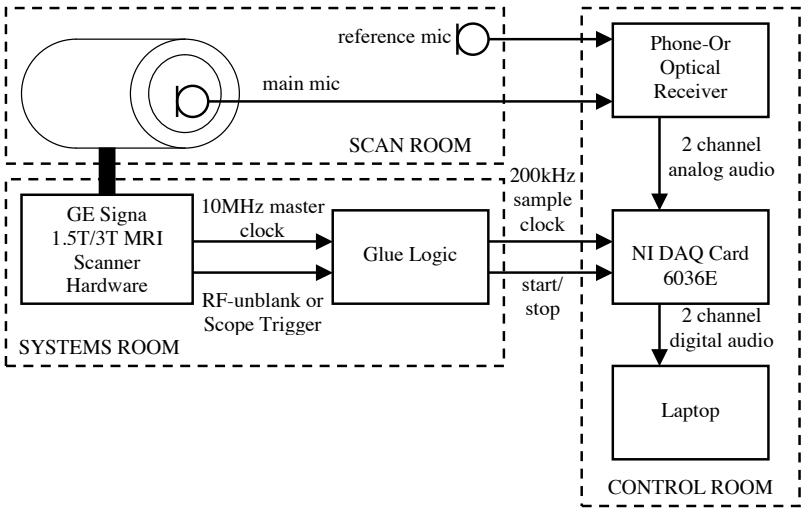


Figure 2.1: System level diagram of the audio acquisition system

The data are recorded on a laptop computer using a National Instruments NI-DAQ 6036E PCMCIA card[2], which provides a total sample rate of 200kHz and supports up to 16 analog input channels. The main and the reference microphone signal are sampled at 100kHz each.

In order to guarantee sample-exact synchronicity the audio sample clock is derived from the MRI scanner's 10MHz master clock. Furthermore, the audio recording is started and stopped using the radio frequency (RF)-unblank signal of the scanner with the help of some interfacing glue logic. This mechanism is described in detail in the following section.

2.4 Synchronizing hardware

The GE Signa scanner provides a digital 10MHz master clock signal to its MRI excitation and readout sequencer circuits, which is also available on the scanner's service interface. Furthermore, the scanner allows access to the digital RF-unblank signal, which is a short

[2] `http://www.ni.com`

low-pulse in the beginning of each MRI acquisition. The key part of the data acquisition system is the FPGA-based digital glue logic that interfaces the MRI scanner hardware to the audio analog to digital converter (ADC) on the NI-DAQ card. The logic circuitry was implemented on a DIGILAB 2 XL board[3], which contains a XILINX Spartan 2 FPGA[4].

The digital glue logic consists of two independent systems, namely a clock divider and a re-triggerable monostable. The clock divider derives a 200kHz clock signal from the 10MHz master clock, which is used to clock the ADC on the NI-DAQ card, resulting in a sampling rate of 100kHz for each of the two microphone channels.

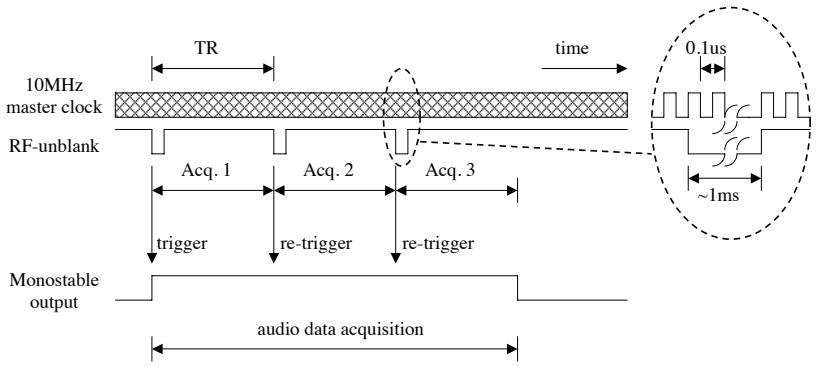


Figure 2.2: Glue logic timing diagram

The re-triggerable monostable vibrator has a time-constant which equals the MRI repetition time (TR). The monostable is (re-)triggered on the falling edge of each RF-unblank pulse, i.e., in the beginning of each MRI acquisition. If a number of MRI acquisitions are performed consecutively a train of RF-unblank low-pulses is observed with a time distance of TR. Each RF-unblank pulse re-triggers the monostable and keeps its output high during the entire acquisition period. This process is shown in Figure 2.2, where we assume a series of three consecutive MRI acquisitions.

[3] http://www.digilentinc.com
[4] http://www.xilinx.com

The output of the monostable is used as an enable signal for the NI-DAQ ADC. This mechanism turns on the analog-to-digital conversion with the first MRI acquisition in a series and stops it as soon as the RF-unblank pulses disappear, i.e. exactly one TR after the last unblank pulse of the acquisition series was observed.

The enable delay of the NI-DAQ card is on the order of 100ns which is negligible with respect to the audio sample time of 50μs at 20kHz. Therefore the audio recording begins almost exactly when the MRI acquisitions start. And since the ADC sample clock is directly derived from the MRI scanner's 10MHz clock signal, which governs the image acquisition, the audio and the MRI images are always exactly synchronized.

2.5 Software components

2.5.1 Data acquisition and sample rate conversion

The real-time data acquisition routine was written in MATLAB[5] and it uses the Data Acquisition Toolbox. In the first post-processing step, low-pass filtering and decimation of the audio data to a sampling frequency of 20kHz is carried out. Finally, the processed audio is merged with the reconstructed MRI image sequence using the VirtualDub software[6].

2.5.2 Noise cancellation

The proposed hardware setup allows for a variety of noise canceling solutions. We describe two noise cancellation methods that we developed: a direct adaptive cancellation method using the well-known normalized least mean square (NLMS) algorithm, and a novel,

[5] `http://www.themathworks.com`
[6] `http://www.virtualdub.org`

model-based adaptive cancellation procedure, which yielded the best results in our speech and singing production experiments.

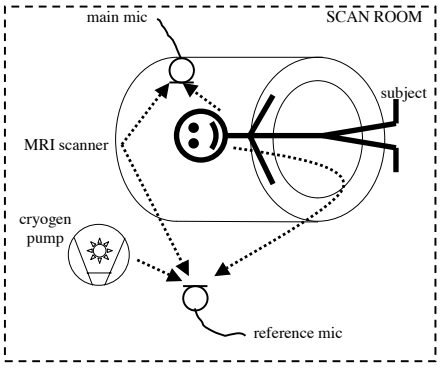


Figure 2.3: Noise sources and microphone arrangement

Figure 2.3 illustrates the location of the microphones and the main sources of noise in the scan room in the proposed set up, namely the subject, the MRI scanner, and the cryogen pump. The dotted lines symbolize the path of the sound, omitting the reflections on the walls of the scan room: The subject's speech is first of all picked up by the main microphone, but there is also a leakage path to the reference microphone. The MRI gradient noise is picked up by both the main and the reference microphone through different paths and hence with different time delays and different filtering, but with similar intensity. Lastly, the cryogen pump noise affects mainly the reference channel.

The GE Signa scanner also has an integrated cooling fan which produces some air flow through the bore of the magnet. The fan may also produce additional noise but can be turned off during the scan. In our experiments, however, we found the fan noise negligible.

It should be noted that the MRI gradient noise is by far the strongest of all noise sources. But despite its high power, it also has some advantageous characteristics, namely it is stationary, periodic and directly dependent on the MRI pulse sequence. In our case we used a 13-interleaf spiral gradient-echo sequence with an echo time (TE) of 0.9ms,

and a TR of 6.856ms, which results in a period of 89.12ms. This means that the scan noise can be thought of as a periodic function with a fundamental frequency of 11.22Hz. As will be shown below, this characteristic can be exploited to achieve very good noise cancellation results within a modeled-reference framework.

2.5.2.1 Direct NLMS noise cancellation

In order to overcome the above-mentioned limitations, a noise cancellation procedure was developed which is based on the well-known NLMS algorithm [23, 28]. The corresponding system diagram is shown in Figure 2.4: The MRI gradient noise $n(t)$ is assumed to be filtered by two independent linear systems H_1 and H_2, which represent the acoustic characteristics of the room, before it enters the main and reference channel microphones, respectively. The speech signal $s(t)$ on the other hand is captured directly by the main channel microphone.

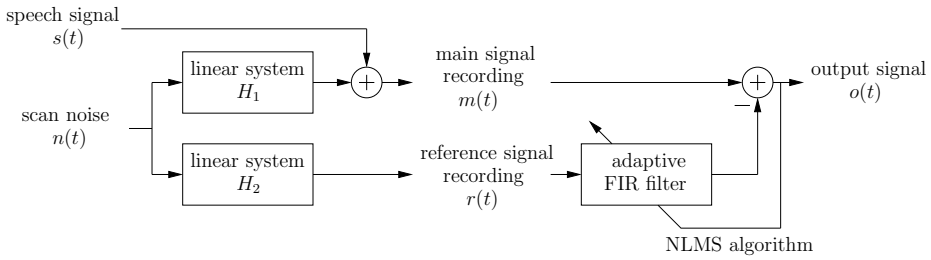


Figure 2.4: Adaptive FIR filter using NLMS algorithm for direct cancellation of interference from MRI scanner noise

During the post-processing, the reference signal $r(t)$ is fed into an adaptive finite impulse response (FIR) filter, and subsequently subtracted from the main channel $m(t)$. The NLMS algorithm continually adjusts the FIR filter coefficients in such a way that the average output signal power is minimized. Or, in other words, the adaptive FIR filter is continuously adjusted in a way that it best approximates the transfer function $\frac{H_1}{H_2}$.

Since the noise cancellation is done off-line in our setup, the FIR filter is allowed to be non-causal and the noise cancellation can be achieved regardless whether the time delay between main and reference channel is positive or negative.

The adaptive FIR filter in our case was of order 4000, and the sampling frequency was 20kHz. The updating coefficient was set to 0.5. The achieved noise reduction was around 17dB. Further details are provided in Section 2.6.

2.5.2.2 Model-based NLMS noise cancellation

A much improved noise reduction was achieved using an artificial reference signal, r_{m}, which is generated based on a pulse-sequence specific model for the MRI gradient noise, rather than the reference signal captured during the scan. The corresponding system diagram is shown in Figure 2.5. Hereby we exploit the periodic nature of the gradient noise, and we generate a signal $r_{\mathrm{m}}(t)$ consisting of the sum of unity-amplitude sinusoids of the fundamental frequency of the MRI scan noise, e.g. $f_1 = \frac{1}{N_{\mathrm{interleaves}} \cdot \mathrm{TR}}$, and all integer multiples up to half the audio sampling frequency. For our settings of TR=6.856ms and $N_{\mathrm{interleaves}} = 13$ interleaves we have $f_1 = 11.22\mathrm{Hz}, f_2 = 22.44\mathrm{Hz}, \ldots, f_{891} = 9996.86\mathrm{Hz}$. Hence, this signal contains all spectral components that the periodic gradient noise waveform can possibly have in the audio frequency band. The signal now serves as the reference for an NLMS noise canceller with an FIR filter of order 4000, with an updating coefficient of 0.5. The achieved noise suppression was around 32dB, and details are provided in Section 2.6.

While the proposed system of Figure 2.5 achieves decent noise suppression results and is straightforward to implement, it should be noted that, as shown in [21], given the sinusoidal nature of the interference the weights of the adaptive FIR filter cannot converge to a constant value and hence the filter update factor cannot be made arbitrarily small.

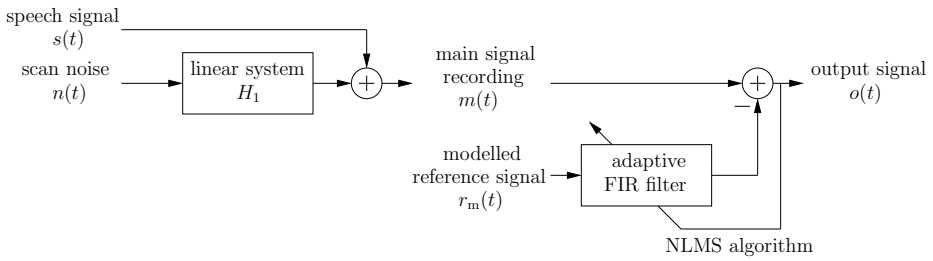


Figure 2.5: Adaptive FIR filter using NLMS algorithm for model-based cancellation of interference from MRI scanner noise

Instead the filter weights must be free to oscillate, though the magnitude of the oscillation decreases with the increased filter order. Furthermore, selecting the filter order is not an easy task, and the value of 4000 was chosen through trial and error.

Both issues can be addressed with a revised system structure which is based on an algorithm described in [65]. The revised system is shown in Figure 2.6, and it contains a bank of complex adaptive filters which are each fed with the sine and cosine signal of the individual harmonics in the noise model. Here the coefficients converge and the filter order is exactly defined by the number of harmonics of the scan noise in the audio band. The achieved noise cancellation result is similar to that of the system shown in Figure 2.4 but the revised system is computationally more efficient since it uses only 1782 instead of 4000 coeficients.

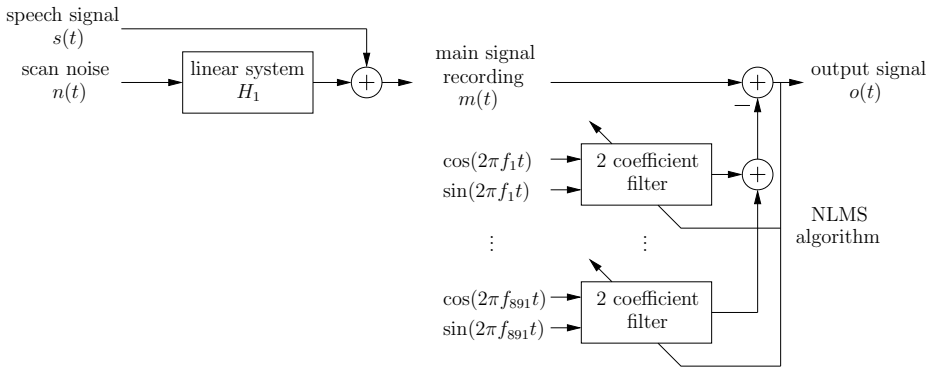


Figure 2.6: Complex adaptive FIR filter using NLMS algorithm for model-based cancellation of interference from MRI scanner noise

Regardless which system structure is used, the disadvantage of the model-based procedures is that they do not account for other noise sources than the MRI scanner, such as the cryogen pump. The major advantage of this approach however is that there is no leakage of the desired signal, i.e. the subject's speech, into the reference channel. Though it might be possible to find a more accurate reference signal model which also includes the cryogen pump, we found that the cancellation of the MRI gradient noise alone provides an output signal with sufficient quality for further analysis.

Another advantage of the model-based procedure is that it lends itself to real-time implementations since even non-causal noise canceling FIR filters are implementable because the modeled reference signal is deterministic.

2.6 Results

In order to quantify the effectiveness of the noise cancellation algorithms a 30 second silence recording was obtained, i.e. without any speech activity, and the average output signal power was measured. Table 2.1 summarizes the achieved noise suppression for unweighted, A-weighted[7], and ITU-R 468 weighted[8] output power measurements.

Table 2.1: Noise power suppression for the two presented methods during no speech

	unweighted (dB)	A-weighted (dB)	ITU-R 468 (dB)
Direct NLMS	17.1	17.6	16.3
Model-based NLMS	32.8	31.1	32.7

The verification of the noise canceller for recordings with speech and/or singing is more difficult since one cannot simply separate the signal and the noise in the recordings and measure their power independently. However, an estimate of the signal to noise ratio (SNR) was obtained by measuring the signal power during speech periods, $P_{\text{speech+noise}}$,

[7]IEC 179 standard available at `http://www.iec.ch/`.
[8]ITU-R 468 standard available at `http://www.itu.int/`.

and scan noise-only periods, P_{noise}, for a given recording. Due to the stationarity of the noise, and the independence of the noise and speech processes, we can compute the signal power as $P_{\text{speech}} = P_{\text{speech+noise}} - P_{\text{noise}}$. The SNR for the given recording can now be expressed as SNR$= \frac{P_{\text{speech}}}{P_{\text{noise}}} = \frac{(P_{\text{speech+noise}} - P_{\text{noise}})}{P_{\text{noise}}}$. This computation was carried out for the original main channel recording, the direct noise-cancelled output, and the model-based noise-cancelled output. The improvements in SNR with respect to the original recording are summarized in Table 2.2. The corresponding signal waveforms are shown in Figure 2.7. Here we see the main channel recording, the directly noise-cancelled output, the model-based noise-cancelled output and the voice activity flag of the sample utterance "We look forward to your abstracts by December 19th. Happy holidays! [singing]."

Table 2.2: Noise power suppression for the two presented methods during speech

	unweighted (dB)	A-weighted (dB)	ITU-R 468 (dB)
Direct NLMS	17.2	18.5	13.3
Model-based NLMS	28.4	29.7	26.5

Furthermore, we observed a slight echo-like artifact in the audio output signal most likely believed to result from the following: After convergence (say in a no-speech period), the adaptive noise canceller acts like a comb filter and effectively nulls out all frequencies that are integer multiples of the gradient noise fundamental. If now suddenly a speech signal appears, which generally has energy at those frequencies, the noise-canceling filter will take some time to adapt and again block out these frequencies. When the speech segment is over, the filter again needs a short time to converge back to the no-speech setting. During this time the audio output obviously contains a residue of the reference signal causing a reverberant effect.

As a possible remedy for this effect, one can make the adaptation of the filter dependent on voice activity, such that during the no-speech phases the filter adapts fast, whereas during speech phases the adaptation is slow, or even turned off completely.

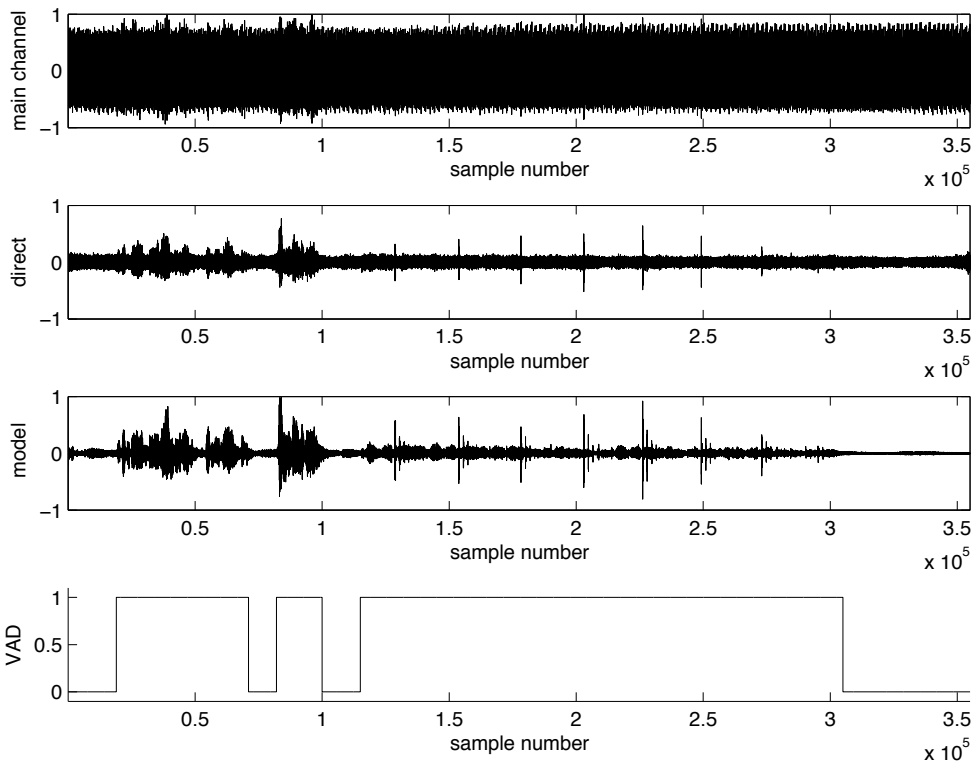


Figure 2.7: Sample waveforms for SNR estimation

Another explanation for the observed residual noise may be that the during the speech production, when the subjects mouth is open and moving, the broad band scan noise excites the resonances of the oral cavity. Some test recordings in which the subject only mouthed an utterance without actually producing the speech contain an obvious sound signature of the utterance used. While at this point no solution to this problem can be proposed it should be noted that even with the currently achieved limited noise cancellation a useful audio signal can be supplied for further analysis in speech production studies.

Chapter 3

MR image processing

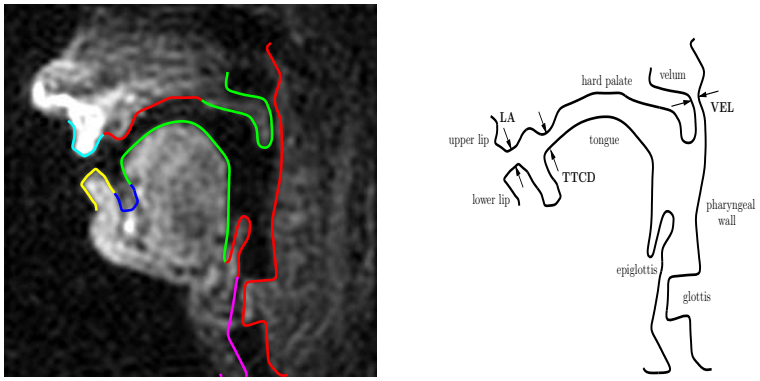
3.1 Abstract

A method is described for unsupervised region segmentation of an image using its spatial frequency domain representation. The algorithm was designed to process large sequences of RT-MR images containing the 2-dimensional midsagittal view of a human vocal tract airway. The segmentation algorithm uses an anatomically informed object model, whose fit to the observed image data is hierarchically optimized using a gradient descent procedure. The goal of the algorithm is to automatically extract the time-varying vocal tract outline and the position of the articulators to facilitate the study of the shaping of the vocal tract during speech production.

3.2 Introduction

The tracking of deformable objects in image sequences has been a topic of intensive research for many years, and many application specific solutions have been proposed. In this part of the dissertation we describe a method which was developed to track tissue structures of the human vocal tract in sequences of midsagittal RT-MR images for the use in speech production research studies. The term "real-time" hereby means that the

MR frame rate is high enough to directly capture the vocal tract shaping events with sufficient temporal resolution, as opposed to the repetition-based cine MRI techniques.



(a) Midsagittal real-time MR image with contours of interest.

(b) Articulators and sample vocal tract variables.

Figure 3.1: Example vocal tract MR image and tract variables

The movie file `movie1.mov`[1], shows an example RT-MR image sequence containing German read speech. This file contains 390 individual images at a rate of approximately 22 frames per second. These MR data were acquired with a GE Signa 1.5T scanner with a custom-made multi-channel upper airway receiver coil. The pulse sequence was a low flip angle 13-interleaf spiral gradient echo saturation recovery pulse sequence with RF and gradient spoiler and slight T1 weighting [45]. The TR was 6.3ms. The reconstruction was carried out using conventional sliding window gridding and inverse Fourier transform operations. Note that a specially designed, highly directional receive coil was necessary to allow for such a small field of view (FOV) without the inevitable MR-typical spatial aliasing, while using spiral read-out trajectories.

A single image has been extracted in Figure 3.1(a) and we have manually traced the contours that are of interest to the speech production researcher. Since the imaging slice is thin (3mm), the air tissue boundaries of the vocal tract tube exhibit a sharp intensity gradient and they appear as distinct edges in the image. Notice that the upper and lower

[1] `http://sail.usc.edu/span/tmi2008/index.php`

front teeth do not show up in the image since their hydrogen content is near zero. The edges which outline the time-varying vocal tract follow these anatomical features:

1. larynx - epiglottis - tongue - lower lip (shown in red),

2. pharyngeal wall - glottis (shown in green),

3. velum - hard palate - upper lip (shown in blue).

These 9 anatomical components, with exception of the hard palate, are called articulators, and they are controlled during the speech production process.

Knowledge about the position and movements of the articulators is fundamental to research on human speech production. More specifically, in the articulatory phonology framework introduced in [50], tract variables are commonly defined to provide a low order description of the shape of the vocal tract at a particular point in time. These variables measure constriction degree and location between various articulators. As an example, Figure 3.1(b) shows some tract variables of interest, such as the lip aperture (LA), the velum aperture (VEL), as well as the tongue tip constriction degree (TTCD), but other tract variables between different pairs of articulators can be defined.

For the time evolution of the vocal tract shaping to be studied in speech production experiments, many such image sequences are acquired from a particular subject, which contain numerous different utterances of interest, and subsequently the relevant tract variables need to be extracted. This task comprises for each image the tracing of the air-tissue boundary of the articulators and the search for the minimum aperture(s) in the appropriate regions.

From an image processing point of view, the problem to be solved is not just that of contour detection but also object identification. This is to say, it is not sufficient to automatically identify in each image the air-tissue boundaries shown in Figure 3.1(a).

Instead, the anatomical subsections of those boundaries corresponding to the articulators have to be, at least approximately, found as well in order to be able to compute the tract variables.

This contour tracing process is time consuming and tedious when carried out by a human. As an example, assuming a time requirement of 3 minutes for the manual tracing of a single image the tracing of the entire 390-frame example image sequence above would take almost 20 hours. Hence, it is the goal of this article to provide an algorithm for the unsupervised extraction of the outline of the individual articulators to facilitate an automatic computation of the tract variables. It is desired to reduce the required human interaction to a minimum, limiting it only to a one-time manual initialization step for a particular subject, whose data will be used for all images of all sequences recorded from said subject.

Since the tracing process can be carried out off-line, i.e., after completion of the MR data acquisition, there is no requirement for real-time execution of the tracing task. However, we desire to process each image of a particular sequence independently, so as to be able to achieve fast tracing of a sequence through the use of parallel image processing on a computing cluster.

Overview and contributions

From an algorithmic standpoint, a main contribution of this paper is a formulation of the edge detection problem in the spatial frequency domain, where we utilize the closed-form solution of the Fourier transform of polygonal shape functions. For the intended application of our method in the context of upper airway MR imaging and vocal tract contour

extraction, we furthermore propose the use of an anatomically informed geometrical object model in conjunction with a corresponding anatomically informed gradient descent procedure to solve the underlying optimization problem.

This part of the thesis is organized as follows: In Section 3.3 we will review some relevant literature and the research context. In Section 3.4 we will then outline a frequency domain-based algorithm, which addresses the edge detection problem through region segmentation. The algorithm requires solving a non-linear least squares optimization problem, which is handled through a gradient descent procedure. This algorithm operates directly on single channel MR data in k-space and it will be validated using a simple MR phantom experiment.

In Section 3.5 we extend and modify our algorithm to process multi-channel in vivo upper airway MR data. The extensions to the algorithm include some data pre-processing steps, the introduction of an anatomically informed 3-region geometrical model of the upper airway, as well as the use of an anatomically informed gradient descent procedure. The modified version of our algorithm will be validated using linguistically informed example images.

Finally, Section 3.6 includes a discussion of the capabilities, advantages, and disadvantages of the algorithm, as well as conclusions and further research suggestions.

3.3 Research context and literature review

3.3.1 The role of MR technology in speech production research

Due to the fact that MR imaging allows to safely and non-invasively observe the entire vocal tract including the deep-seated structures, this technology has gained much importance in the field of speech research. However, compared to other currently used

modalities such as x-ray [19], ultrasound [64], or electromagnetic articulography [48], MR imaging yields only moderate data rates. Due to the low spatio-temporal resolution of conventional MRI acquisition techniques, the earliest MR-based speech studies were limited to vocal productions with static postures such as vowel sounds (see [45] and the references therein). The subsequent development of cine MR imaging techniques [56][62] allowed the imaging of dynamic vocal tract shaping with sufficient spatial and temporal resolution but this method relies on multiple exact repetitions of the utterance to be studied with respect to a trigger signal. Hence, cine MR imaging may be difficult to use for the study of continuous running speech.

Recent advances in MR pulse-sequence design reported in [45] allow real-time MR imaging of the speech production process at a suitably high frame rate. At the same time, a new image processing challenge has been posed by the necessity of the contour extraction from the real-time MR images which are, generally speaking, of poor quality in terms of noise. A similar problem has been addressed in [19] for the case of x-ray image sequences showing the sagittal view of the human vocal tract. However, midsagittal MR images and sagittal x-ray images are quite different since the x-ray process only allows a projection through the volume of interest, i.e., the head of the subject, so that for instance, the teeth obstruct the view of the tongue. The MR imaging process on the other hand allows the capture of a thin midsagittal slice and hence the contours of interest are not compromised. The algorithm presented in this paper is particularly geared towards an application in the MR imaging context.

3.3.2 Various approaches to edge detection and contour tracking and their methodologies

3.3.2.1 Open versus closed contours

From the problem definition in Section 3.2, it is clear that our task is to identify in each image the air tissue boundary of the articulators, which are open contours. That requires first of all finding the start and end points of the boundary sections. Unfortunately, we have no artificial markers or anatomical landmarks available that can be easily registered. We have previously attempted to solve this problem using the optical flow approach, as reported in [3]. Here the first frame of a given image sequence was manually initialized, and the start and end point locations of the boundary segments were then consecutively estimated for the subsequent images. The main problem with this approach is that any estimation error propagates forward, requiring frequent and time-consuming manual correction throughout the entire contour detection process of an image sequence.

In contrast, a closed contour processing framework appears to be much more attractive, since it is area-based and would be expected to be more noise robust. The authors of [18] have proposed a powerful algorithm to segment from an image a single region with a constant level of intensity, and hence detect this region's boundary against the background. However, this algorithm has two short comings, namely it cannot associate sections of the boundary with certain image features, i.e., anatomical components of interest, and it is only defined for one region of interest. Nevertheless, this procedure inspired our algorithm, and we will cast our problem of identifying the vocal tract articulator boundaries into a multi-area closed-contour framework.

3.3.2.2 Contour descriptors

A central issue in any type of contour tracking or edge detection application is that of contour representation. In general, an ideal boundary contour descriptor allows to express the contours with few parameters and does not produce self-intersecting curves. Additionally, the roughness of the boundary needs to be somehow controllable in order to mitigate the effect of image noise. However, anatomical objects are often smooth but globally enforcing smoothness on an anatomical boundary may result in significant errors if pointy structures, e.g., the velum or the epiglottis in the airway, make up a part of the boundary. We hence wish to have easy local control over the smoothness. That, however, is only useful if we can also robustly identify those sections of the boundary that differ in their local smoothness properties in order to select the local smoothness constraints appropriately.

A variety of contour descriptors have been used for both open and closed 2-dimensional boundaries such as the B-spline [18], wavelet [24], Fourier [52], and polyline descriptors [30], none of which is guaranteed self-intersection free. In our algorithm we use the polyline contour descriptor since it is the only one which affords us a convenient closed form solution of the external energy functional and its gradient with respect to the contour parameters if used for closed polyline contours.

3.3.2.3 Energy functionals

No matter which particular contour detection method is used, in order to evaluate the goodness of the fit of a candidate contour to the underlying observed image a measure E_{ext} is commonly designed. This measure is referred to as external energy. Depending on the application, this measure quantifies how well the contour corresponds to edges in the image, intensity extrema, or other desired image features [40].

In the edge detection case, the external energy is commonly chosen as the line integral along the boundary contour C over the negative image intensity gradient magnitude

$$E_{\text{ext}}(C) = -\int_C |\nabla(h(x,y) * m(x,y))|\, ds \tag{3.1}$$

where $m(x,y)$ is the observed image intensity function, and $h(x,y)$ is an optional smoothing filter kernel, which may be used to mitigate the effects of noise in the image.

This way, if the candidate contour coincides well with an edge in the image, i.e., with a line of high intensity gradient magnitude, the external energy will have a large negative value. The edge finding process now consists of minimizing the external energy by adjusting the shape of the contour through an appropriate optimization algorithm:

$$\hat{C} = \underset{C}{\operatorname{argmin}}\, E_{\text{ext}}(C) \tag{3.2}$$

At this point we can identify a number of mathematical issues: First, the external energy is an integral quantity along the candidate contour, which can be difficult to evaluate analytically depending on the underlying $m(x,y)$ and the shape of the contour. In practice, approximations are often used to resolve this problem. Second, a derivative of $E_{\text{ext}}(C)$ with respect to the parameters of the contour C will be needed if the optimization is to be carried out through a gradient descent procedure, which in practice can oftentimes only be approximated using a finite difference. And third, the convexity of the function $E_{\text{ext}}(C)$ cannot necessarily be guaranteed and a direct gradient descent optimization procedure may get stuck in a local minimum unless a careful initialization of the contour near the optimum location can be carried out.

Furthermore, we identify additional practical problems. First, if the image $m(x,y)$ is noisy in the area of the edge of interest the optimum contour will deviate from the

true underlying edge. If the true edge is known to be smooth, one can combat this effect by designing a penalty term for the optimization problem referred to as internal energy $E_{\text{int}}(C)$. It measures the curvature of the candidate contour (see [30, 40] for detailed information). The optimization problem of Equation 3.2 now becomes

$$\hat{C} = \underset{C}{\operatorname{argmin}} \left(E_{\text{ext}}(C) + E_{\text{int}}(C) \right) \tag{3.3}$$

but the mathematical issues identified above have now become even more difficult to deal with. Secondly, the application of the filter $h(x, y)$ on the image data may smooth away some of the image noise and remove some local minima in the energy landscape of the optimization problem, yet at the same time it may destroy important detail information in $m(x, y)$ and actually hinder the edge finding process.

3.3.2.4 Edge detection in image domain versus frequency domain

Most edge detection algorithms devised to date operate directly on the pixelized image, one exception being the method introduced in [52]. Here, a Fourier contour descriptor is used to capture in a given scene the outline of a single region of interest. The region is modelled to have unity amplitude, and a discretized spatial frequency domain representation of this object model is computed. A nonlinear optimization is subsequently carried out, which aims at matching the frequency domain representation of the model to the frequency domain representation of the observed image by adjusting the boundary contour of the region of interest in the image model.

This approach has the advantage that the accuracy of the boundary is not limited by pixelization in the image space. However, for the particular choice of the Fourier contour descriptor a closed form solution is available neither for the external energy functional

nor for its gradient. The authors in [52] furthermore report a problematically unstable behavior of the optimization algorithm.

We can further note three more shortcomings of this approach, namely, the algorithm is only defined for a single region of interest, second the region's true amplitude in the image may not be unity and there is a mismatch between model and observation, and lastly, the procedure does not match sections of the boundary to particular scene features.

3.3.2.5 Probabilistic approaches to edge detection and tracking

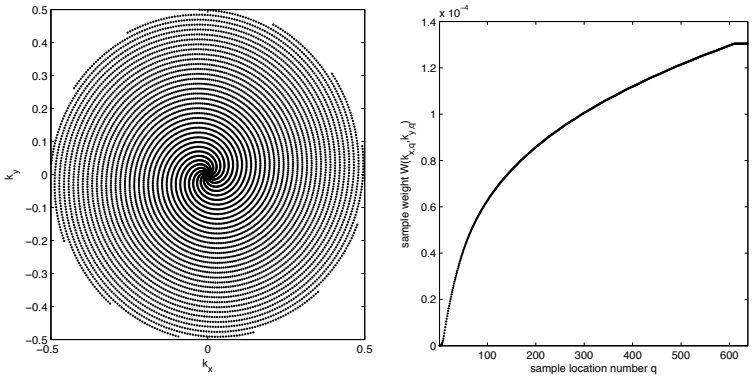
As outlined in [40] and references therein, deformable object models have been used in a probabilistic setting to accomplish the task of edge detection. Here prior knowledge of the probability associated with the object's possible deformation states is required. However, while some statistical models exist for describing vocal tract dynamics during speech, such as the one introduced in [34], the development of comprehensive models that describe natural spontaneous speech phenomena is a topic of ongoing research. Thus, the method presented in this article aims to provide the means to utilize real-time MR imaging technology to produce large amounts of articulatory data for the future development and training of such advanced statistical models.

3.4 Segmentation of MR data in the frequency domain

3.4.1 The concept

The 2-dimensional MR imaging process produces, per frame, Q samples $M(k_{x,q}, k_{y,q}), q = 1 \dots Q$ of the Fourier transform of the spatially continuous magnetization function $m(x, y)$ in the selected imaging slice. Due to non-ideal nature of the MR scanner's receiver coil and circuitry, as well as off-resonance, the samples $M(k_{x,q}, k_{y,q})$ can also be subject to an additional constant phase shift, which in general is not known to the user. In our

case we use a 13-interleaf spiral readout pattern as shown in Figure 3.2(a). It consists of 13 identical (but individually rotated) spiral readout paths which start at the origin of the spatial frequency domain, which is also called k-space. Each spiral readout produces 638 k-space samples, totalling 8294 samples for the k-space coverage as shown. While the MR operator has the capability of applying a zoom factor to increase or reduce the FOV by scaling the readout pattern, the relative geometry of the pattern is the same for all experiments in this article. Notice also that the required number of pixels for the conventional gridded reconstruction process, and with it the relative spatial resolution, is constant regardless of the applied zoom factor, since a scaling of the readout pattern changes equally the radial gap and the k-space coverage area. For the readout pattern used here the image matrix has the size 68-by-68 pixels and all geometrical measurements that appear in this text have been converted into the pixel unit, i.e., any additional FOV change has been accounted for in any of the data below.



(a) Normalized k-space sampling pattern showing all 8294 sample locations.

(b) K-space sample weighting coefficients for a single trajectory.

Figure 3.2: MR readout using 13-interleaf spiral trajectories.

By using the k-space sample values as the coefficients of a truncated two-dimensional Fourier series one can reconstruct an approximation of the underlying continuous magnetization function

$$m(x,y) = \sum_{q=1}^{Q} W(k_{x,q}, k_{y,q}) M(k_{x,q}, k_{y,q}) e^{j2\pi(k_{x,q}x + k_{y,q}y)}, \tag{3.4}$$

where $W(k_{x,q}, k_{y,q})$ are the read-out trajectory-specific density compensation coefficients. These weighting coefficients are commonly obtained using a Voronoi tessellation of the sampling pattern, and they represent the area of the Voronoi cell corresponding to each sample point. Figure 3.2(b) shows the weighting coefficients for the 638 samples of one spiral starting at the origin of k-space. Due to the symmetry of the read-out pattern the weighting coefficients are the same for all 13 spiral trajectories.

While $m(x,y)$ is a continuous function it is commonly evaluated only punctually, i.e., sampled on a cartesian grid, and the resulting sample values are displayed as square patches in a pixelized image matrix. While oversampling and interpolation are viable options to increase the fidelity of the patched approximation it is the goal of our approach to circumvent this pixelization process altogether and carry out the contour finding on the continuous function $m(x,y)$ directly. As benefits of this methodology we expect an improved edge detection accuracy as well as simple and straightforward methods to mathematically evaluate the external energy of the contour and its gradient.

Figure 3.3 shows a flow chart of our proposed algorithm using some sample upper airway images, which for the specified FOV always consist mainly of three large connected regions of tissue. The MR data acquisition (left hand side in the figure) produces k-space samples of the object which are collected in the Q-by-1 column vector $\mathbf{M}$. On the right hand side in the figure, we have a geometrical object model that consists of three disjoint regions $R_{1\ldots3}$ in the image domain, described by their polyline boundaries and

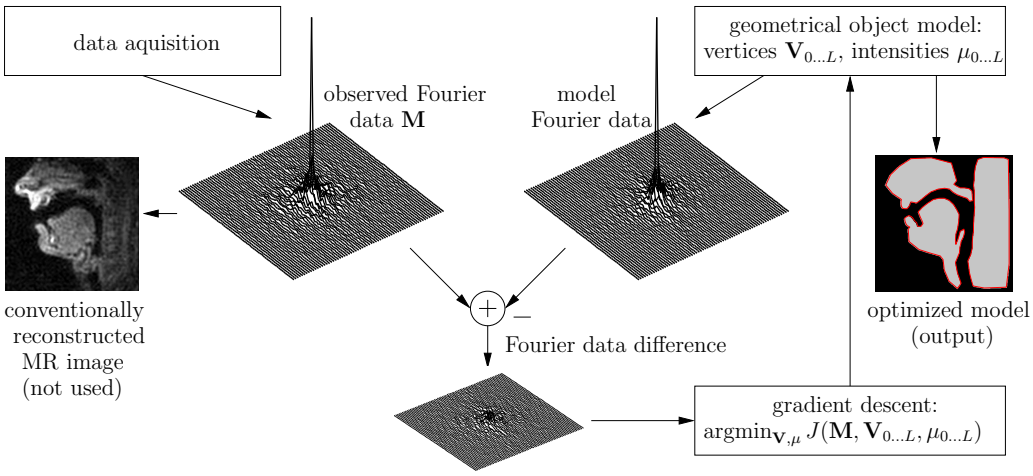


Figure 3.3: Overview of Fourier domain region segmentation.

their intensities $\mu_{1...3}$ (see also Figure 3.10(a) and Figure 3.10(b)). These three regions are additive to a square background R_0 with intensity μ_0, which spans the entire FOV. The parameters of our geometrical model are the vertex vectors of the polylines $\mathbf{V}_{1...3}$, as defined in Equation A.1 as well as the intensities $\mu_{0...3}$. From these geometrical parameters, we derive the frequency domain representation of the model with help of the analytical solution of the 2-dimensional Fourier transform of the polygonal shape functions. Details of this key mathematical component of our algorithm can be found in the Appendix. An optimization algorithm is then used to minimize the mean squared difference $J(\mathbf{M}, \mathbf{V}_{1...3}, \mu_{0...3})$ between the observed frequency domain data and the frequency domain data obtained from the model by adjusting the model's parameters and hence improving the model's fit to the observed scene. Hereby, the model's frequency domain representation is obtained at the same k-space sampling locations that were used for the observed data. Hence, any spatial aliasing and/or Gibbs ringing affects the model in the same way as the object.

The final desired output of the algorithm for a given set of input data is the geometrical description of the best-fitted polyline boundary of the model. Using the polyline

boundaries one can then easily compute the apertures that correspond to the vocal tract variables.

At this point it must be clearly stated that the geometrical model must be able to sufficiently approximate the underlying observed data. If any other FOV is chosen, or even a totally different object is imaged then the geometrical model has to be re-designed accordingly.

3.4.2 The mathematical procedure

To rigorously derive our algorithm we start in the continuous spatial domain with an object model consisting of $L = 3$ disjoint regions plus the background. We express the difference image energy between the underlying observed function $m(x, y)$ and the model as

$$J = \iint \left| m(x,y) - \sum_{l=0}^{L} \mu_l s(x, y, \mathbf{V}_l) \right|^2 dxdy \tag{3.5}$$

where $s(x, y, \mathbf{V}_l)$ is a polygonal shape function as defined in Equation A.2.

Using Parseval's Theorem and Equation 3.4 we now step into the frequency domain by sampling the frequency domain representation of the model on the same grid as the measured frequency domain data. With $S(k_x, k_y, \mathbf{V}_l)$ being the Fourier transform of $s(x, y, \mathbf{V}_l)$ as stated in Equations A.4 and A.5 we write

$$J\left(\mathbf{M}, \mathbf{V}_{0\ldots L}, \mu_{0\ldots L}\right) = \sum_{q=1}^{Q} W(k_{x,q},k_{y,q}) \left| M(k_{x,q}, k_{y,q}) - \sum_{l=0}^{L} \mu_l S(k_{x,q}, k_{y,q}, \mathbf{V}_l) \right|^2 \tag{3.6}$$

To express this in matrix form, we define

$$\bar{\mathbf{M}} = \left[\sqrt{W(k_{x,1}, k_{y,1})} M(k_{x,1}, k_{y,1}), \ldots, \sqrt{W(k_{x,Q}, k_{y,Q})} M(k_{x,Q}, k_{y,Q}) \right]^T \tag{3.7}$$

which is a Q-by-1 column vector consisting of the observed frequency domain measurements multiplied by their corresponding weighting factor, and

$$\bar{\mathbf{S}}(\mathbf{V}_l) = \left[\sqrt{W(k_{x,1},k_{y,1})}S(k_{x,1},k_{y,1},\mathbf{V}_l),\ldots,\sqrt{W(k_{x,Q},k_{y,Q})}S(k_{x,Q},k_{y,Q},\mathbf{V}_l)\right]^T \quad (3.8)$$

which is a Q-by-1 vector of the frequency domain representation of the region l at the same spatial frequencies as the observed data, multiplied by the same weighting factor. Furthermore, we define

$$\boldsymbol{\Psi}(\mathbf{V}_{0\ldots L}) = [\bar{\mathbf{S}}(\mathbf{V}_0),\ldots,\bar{\mathbf{S}}(\mathbf{V}_L)] \quad (3.9)$$

which is a Q-by-$(L+1)$ matrix combining the frequency domain representations of the L segments and the background, and

$$\boldsymbol{\mu} = [\mu_0,\ldots,\mu_L]^T \quad (3.10)$$

which is a $(L+1)$-by-1 vector containing the intensities of the L regions and the background. Lastly, we write the scalar objective function J as

$$J\left(\bar{\mathbf{M}},\mathbf{V}_{0\ldots L},\boldsymbol{\mu}\right) = \left\|\bar{\mathbf{M}} - \boldsymbol{\Psi}(\mathbf{V}_{0\ldots L})\boldsymbol{\mu}\right\|^2 \quad (3.11)$$

The goal is now to minimize

$$\hat{\mathbf{V}}_{1\ldots L},\hat{\boldsymbol{\mu}} = \underset{\mathbf{V}_{1\ldots L},\boldsymbol{\mu}}{\operatorname{argmin}} J\left(\bar{\mathbf{M}},\mathbf{V}_{0\ldots L},\boldsymbol{\mu}\right) \quad (3.12)$$

in order to find the best fitting model to the observed data, i.e., we optimize the boundary curves of the L segments (but not the background) as well as all intensities.

Estimation of the region intensities

Following [22], we separate from the optimization problem posed in Equation 3.12 the linear part, and for a given model region geometry $\mathbf{V}_{0\ldots L}$ we jointly estimate the intensities of all model regions including the background as

$$\hat{\boldsymbol{\mu}} = \underset{\boldsymbol{\mu}}{\operatorname{argmin}} \left\| \bar{\mathbf{M}} - \boldsymbol{\Psi}(\mathbf{V}_{0\ldots L})\boldsymbol{\mu} \right\|^2 \tag{3.13}$$

We can solve this system of equations in the minimum mean square sense with the pseudo-inverse $\boldsymbol{\Psi}^+ = \left(\boldsymbol{\Psi}^H \boldsymbol{\Psi}\right)^{-1} \boldsymbol{\Psi}^H$ and we obtain

$$\hat{\boldsymbol{\mu}} = \boldsymbol{\Psi}^+(\mathbf{V}_{0\ldots L})\bar{\mathbf{M}} \tag{3.14}$$

Notice at this point that the region intensities $\hat{\boldsymbol{\mu}}$ can obtain complex values.

Estimation of the region shapes

The result of Equation 3.14 can be used to simplify Equation 3.11 to

$$J\left(\bar{\mathbf{M}}, \mathbf{V}_{0\ldots L}\right) = \left\| \bar{\mathbf{M}} - \boldsymbol{\Psi}(\mathbf{V}_{0\ldots L})\boldsymbol{\Psi}^+(\mathbf{V}_{0\ldots L})\bar{\mathbf{M}} \right\|^2 \tag{3.15}$$

and we now state the optimization goal for $\mathbf{V}_i, i = 1 \ldots L$

$$\hat{\mathbf{V}}_i = \underset{\mathbf{V}_i}{\operatorname{argmin}}\, J\left(\bar{\mathbf{M}}, \mathbf{V}_{0\ldots L}\right) \tag{3.16}$$

We will tackle this unconstrained non-linear optimization problem using a gradient descent procedure. Let $\mathbf{v}_{i_j}$ be a vertex vector belonging to the boundary polyline $\mathbf{V}_i$

of region R_i. The gradient of the objective function J with respect to this vector's x-coordinate is

$$\frac{\partial J}{\partial x_{i_j}} = -2\Re\left\{\left(\bar{\mathbf{M}} - \boldsymbol{\Psi}\hat{\boldsymbol{\mu}}\right)^H \left(\frac{\partial \boldsymbol{\Psi}}{\partial x_{i_j}}\hat{\boldsymbol{\mu}} + \boldsymbol{\Psi}\frac{\partial \hat{\boldsymbol{\mu}}}{\partial x_{i_j}}\right)\right\} \tag{3.17}$$

In Equation 3.17, the derivative $\frac{\partial \boldsymbol{\Psi}}{\partial x_{i_j}}$ expresses how the model's frequency domain representation changes if the x-coordinate of $\mathbf{v}_{i_j}$ is changed. This term can be obtained using Equation A.12 from the Appendix. With Equation 3.14 the term $\boldsymbol{\Psi}\frac{\partial \hat{\boldsymbol{\mu}}}{\partial x_{i_j}}$, can be written as

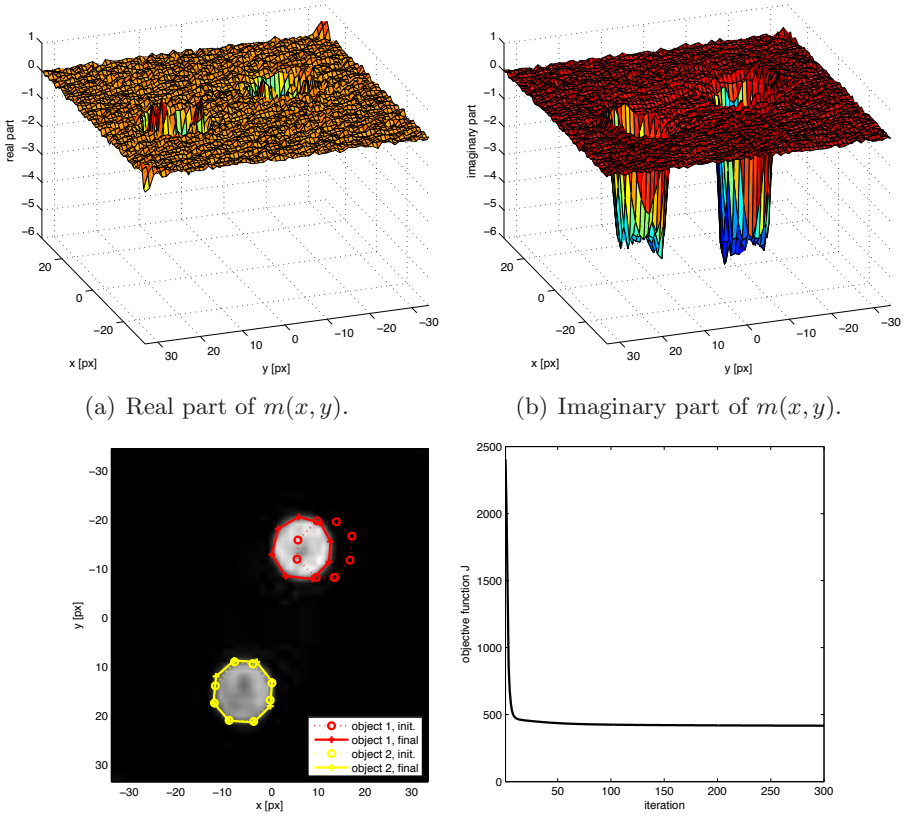
$$\boldsymbol{\Psi}\frac{\partial \hat{\boldsymbol{\mu}}}{\partial x_{i_j}} = \boldsymbol{\Psi}^{+^H}\frac{\partial \boldsymbol{\Psi}^H}{\partial x_{i_j}}(\bar{\mathbf{M}} - \boldsymbol{\Psi}\hat{\boldsymbol{\mu}}) - \boldsymbol{\Psi}\boldsymbol{\Psi}^+\frac{\partial \boldsymbol{\Psi}}{\partial x_{i_j}}\hat{\boldsymbol{\mu}} \tag{3.18}$$

where we used the formula for the derivative of an inverse matrix[2] $\frac{d\boldsymbol{A}^{-1}}{dt} = -\boldsymbol{A}^{-1}\frac{d\boldsymbol{A}}{dt}\boldsymbol{A}^{-1}$. Similarly, the gradient of J with respect to the vertices y-coordinate can be found along the same lines, and both can be combined into the vector $\frac{\partial J}{\partial \mathbf{v}_{i_j}} = \left[\frac{\partial J}{\partial x_{i_j}}, \frac{\partial J}{\partial y_{i_j}}\right]$.

At this point we would like to carry out a simple gradient descent for each vertex vector in the model

$$\mathbf{v}_{i_j}^{(n+1)} = \mathbf{v}_{i_j}^{(n)} - \epsilon\frac{\partial J}{\partial \mathbf{v}_{i_j}}^{(n)} \tag{3.19}$$

but the convexity of the objective function J cannot be guaranteed. Hence for all but the simplest object geometries and a close initialization of the model to the true edge location the descent is highly likely to get stuck in a local minimum. To overcome this problem for the upper airway scenario we propose in Section 3.5 the use of an anatomically informed object model in conjunction with a hierarchical optimization algorithm. However, for a first validation of our method we will now present the results of a simple phantom experiment which utilizes the direct gradient descent optimization.



(a) Real part of $m(x, y)$.

(b) Imaginary part of $m(x, y)$.

(c) Magnitude image $|m(x, y)|$, initial contours, final contours.

(d) Time course of the objective function.

Figure 3.4: Phantom experiment data, initial contours, and final contours.

3.4.3 An experiment with a simple stationary phantom

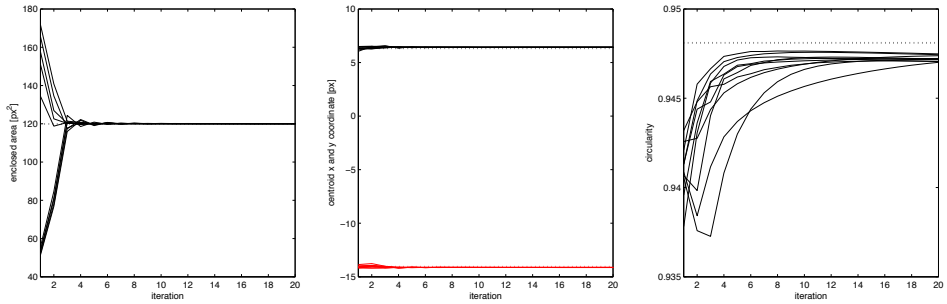
For our first edge detection validation experiment we imaged the cross-sectional slice of two stationary cylindrical objects, which were thin-walled plastic containers filled with butter[3] using a GE Signa 1.5T scanner with single channel bird-cage head coil. A conventionally reconstructed magnitude image of the phantom is shown in Figure 3.4(c), and the goal of the first experiment is to determine the location of the boundaries of the two objects. Notice that the rasterized approximation of $|m(x, y)|$ is displayed in

[2]http://planetmath.org/encyclopedia/DerivativeOfInverseMatrix.html

[3]Water was found not suitable as phantom material since it warms up during the scan process and flow-artifacts appear in the image. Hence the containers were filled with butter instead, then heated and subsequently cooled, so as to achieve a complete filling of the volume with solid material.

Figure 3.4(c) for illustration purposes only, while the segmentation algorithm utilizes the samples $M(k_{x,q}, k_{y,q})$ directly, whose corresponding complex valued transform function $m(x, y)$ is shown in Figure 3.4(a) (real part) and Figure 3.4(b) (imaginary part).

The object model for this experiment consisted of the background and two additive octagons, one for each of the two circular phantom areas. While we carefully initialized the boundary for one of the circular regions (object 2) we deliberately mis-initialized the other (object 1) to be significantly offset. The initial boundaries are shown as dotted lines in Figure 3.4(c) and polyline vertices are shown with "o" markers, while the final boundaries are shown as solid lines with "+" markers. Figure 3.4(d) shows how the objective function J converges over the course of 300 direct gradient descent steps with a step width $\epsilon = 0.01$. And we see that eventually both objects' boundaries are well captured by the algorithm despite the grossly mis-initialized boundary for object 1.



(a) Enclosed area, and the true enclosed area (dotted).

(b) Centroid coordinates (x- black, y-red).

(c) Circularity, and the maximum possible value (dotted)

Figure 3.5: Edge detection on object 1: time courses of the geometrical accuracy measures for 10 different initializations (initial 30 iterations shown).

Since the true boundaries of the objects are circular whereas the model's boundaries are octagons it is not directly possible to quantify the accuracy of the final achieved boundary detection result. Furthermore, the true locations of the objects are unknown, so we cannot easily evaluate the positional accuracy of our algorithm either. Instead we will compare the enclosed area of the boundary polygon 1 to the true area of the

circular object, which has been determined through 10 manual measurements of the phantom. Furthermore, 10 manual tracings of object 1 have been carefully carried out on the image and the average centroid will be used as the as a substitute for the true center location of the object. This value will then be compared with an averaged centroid from 10 differently initialized automatic edge finding results. Lastly, the averaged achieved circularity[4] measure of the 10 manual tracings and 10 automatic detection results can be compared to the maximum possible value, which for a regular octagon equals $\frac{\pi}{8}(1+\sqrt{2}) \approx 0.9481$.

Table 3.1: Edge detection on object 1: averaged geometrical accuracy measures and standard deviations over 10 trials

	manual (σ)	automatic (σ)	true (σ)
area [px^2]	108.6 (7.1)	119.8 (0.03)	119.9 (0.18)
centroid x [px]	6.34 (0.14)	6.42 (0.0004)	unknown
centroid y [px]	-14.06 (0.10)	-14.13 (0.0022)	unknown
circularity	0.941 (0.0015)	0.945 (0.0002)	0.9481

The averaged results of the 10 validation runs (300 iterations each, $\epsilon = 0.01$) are summarized in Table 3.1, and the initial 30 values of the time courses of the various measures are shown in Figure 3.5(a) and Figure 3.5(b), and 3.5(c). For the enclosed polygon area we find that the automatic detection results are well within one standard deviation of the true value of 119.9 square pixels, whereas the manual tracings were on average significantly too small. The centroid x and y coordinates of manual and automatic tracing coincide well at around $(6.4, -14.1)$, though the manual results exhibit larger variations than the automatic tracking results. It is also important to note that the standard deviations of both the manual and automatic processing are in the sub-pixel range, i.e. both human as well as automatic tracing results in sub-pixel spatial accuracy. Finally, the circularity measure, which equals 1 for a perfect circle and 0.9481

[4]The circularity of a polygon is defined as $4\pi\frac{\text{area}}{\text{perimeter}^2}$. It is a dimensionless quantity whose maximum value equals 1, which is achieved for a perfectly circular area.

for a perfectly regular octagon, shows that on average the automatic tracings are closer to the optimum value than the manual tracings.

In order to investigate the dependency of the algorithm on the completeness of the geometrical model we repeat the experiment with the same parameters but with an incomplete geometrical model which only includes a region for object 1 in addition to the background. The intial and final boundaries for object 1 are shown in Figure 3.6 superimposed on the magnitude image, and we observe that the shape and the location of object 1 are again well captured despite the obvious mismatch between geometrical model and observed scene.

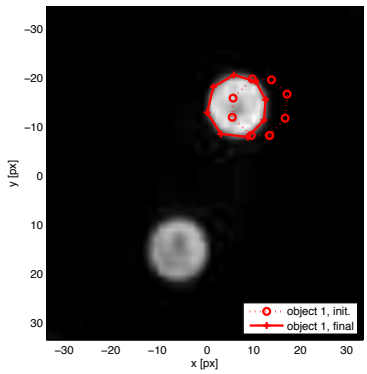


Figure 3.6: Edge detection on object 1 with single-region geometrical model: Magnitude image $|m(x, y)|$, initial contours, final contours.

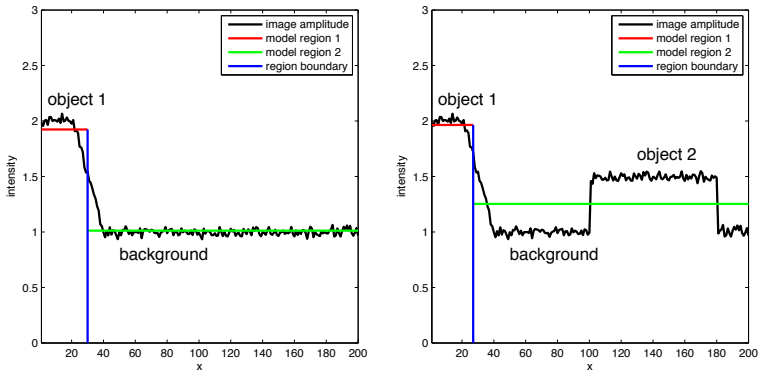
Table 3.2 summarizes the averaged measurements and standard deviations of enclosed area, centroid coordinates, and circularity of 10 trials with different initial contours. While the centroid coordinates and the circularity value are virtually equivalent to those achieved with a complete geometrical model (Table 3.1, column 3) the enclosed area was estimated smaller than previously, yet still more accurately than the manual results.

The reason for this outcome is illustrated in Figure 3.7 with the help of an example. In Figure 3.7(a) we consider the noisy image intensity profile (black line) obtained at a particular y-coordinate of a hypothetical scene which includes an object (on the left;

Table 3.2: Edge detection on object 1 with incomplete geometrical model: averaged geometrical accuracy measures and standard deviations over 10 trials

	automatic (σ)
area [px^2]	119.1 (0.03)
centroid x [px]	6.42 (0.0004)
centroid y [px]	-14.13 (0.0022)
circularity	0.945 (0.0002)

mean amplitude 2) and the background (on the right; mean amplitude 1). The geometrical model consists only of a single constant-intensity region (red) to capture the object in addition to the background (green). The optimum boundary (blue) for this scenario is found at $x = 30$. If, on the other hand, an additional object 2 is present in the scene (Figure 3.7(b); mean amplitude 1.5) but the geometrical model does not account for it then the optimum boundary is found at $x = 27$ since the background amplitude is estimated larger. However, the shift of the boundary location would not occur if the transition between object 1 and the background amplitude were very steep.



(a) Complete geometrical model: boundary location at $x = 30$.

(b) Incomplete geometrical model: boundary location at $x = 27$.

Figure 3.7: Edge location dependency on completeness of the geometrical model.

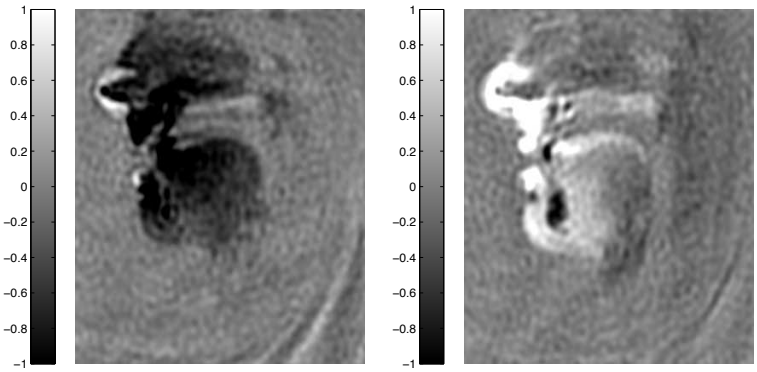
In summary we conclude from this simple phantom experiment that our region-based contour detection algorithm works well for objects with a simple geometry and sharp

boundaries, and it meets or surpasses human contour tracing performance. We do require, however, a good geometrical model which is capable of capturing all objects in the scene.

In the following section we will address the problem of multi-coil human upper airway region segmentation.

3.5 Upper airway multi-coil MR data segmentation

3.5.1 Data pre-processing



(a) Coil 1: real part. (b) Coil 1: imaginary part.

(c) Coil 2: real part. (d) Coil 2: imaginary part.

Figure 3.8: Multi-coil in vivo upper airway MR sample images.

In order to capture the entire human vocal tract a specialized highly directional multi-channel MR receiver coil was employed, and we utilized data from two coils that were

located in front of the subject's face. For a sample image, the complex valued magnetization function is shown Figure 3.8, where Figure 3.8(a) and Figure 3.8(b) show the real and imaginary part reconstructed from coil 1, and Figure 3.8(c) and Figure 3.8(d) those from coil 2. In these images light intensity means positive values, whereas dark intensity means negative values. These images clearly demonstrate that each coil has its own (unknown) spatially varying phase offset, and it is not clear at this point what the best strategy might be to combine the two coils' complex data.

From the magnitude images produced from each coil (Figure 3.9(a) and Figure 3.9(b)) we additionally conclude that we cannot use a single constant-amplitude 3-region object model since the 3 regions of interest (Figure 3.1(a)) show up only partly in each of the coils' data. This effect is due to the spatial roll-off of the individual coil sensitivity functions.

We hence proceed at this point with conventionally reconstructed 68-by-68 cartesian sampled root sum square (RSS) magnitude images such as the one shown in Figure 3.9(c), and we apply the thin-plate spline-based intensity correction procedure proposed in [38] to obtain an estimate of the combined coil sensitivity map (Figure 3.9(d)), which is constant for all images contained in the sequence. We can thus obtain corrected maximally flat magnitude images showing 3 connected constant-amplitude regions of tissue (Figure 3.9(e)). These images' cartesian 68-by-68 Fourier transform matrices will be the starting point for the area-based contour detection following the mathematical procedure as outlined in Section 3.4.

While we have now stepped away from using the MR data directly in k-space, we continue to use our frequency domain based segmentation framework since it affords us a very convenient way to compute in closed form the external energy of the contours as well as the corresponding gradient without any additional interpolation or zero-padding

(a) Coil 1 magnitude.

(b) Coil 2 magnitude.

(c) RSS image.

(d) Estimated combined coil sensitivity pattern.

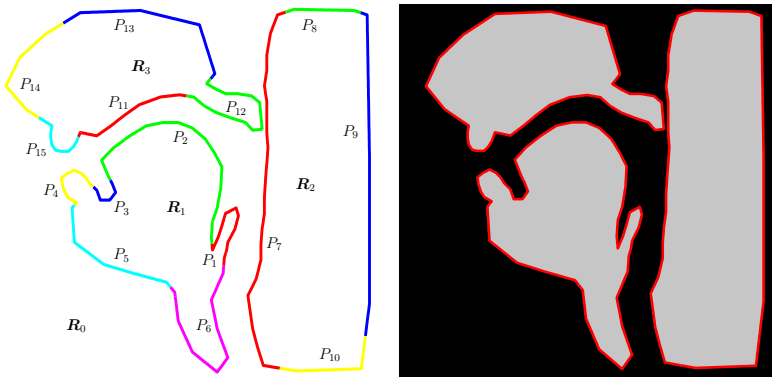
(e) Sensitivity corrected RSS image.

Figure 3.9: Multi-coil upper airway images.

operations. In fact, it is clear that processing in the discretized Fourier domain, given a sufficiently densely sampled Fourier representation of the image, is equivalent to operating on an infinitely accurately sinc-interpolated spatial representation of the image. This fact

is easy to understand if one considers that a perfect sinc-interpolation of an image is just the same as infinite zero-padding in the Fourier domain. That means that processing on the original non-zeropadded Fourier domain samples exploits the same information as processing a perfectly sinc-interpolated image, as derived using Equations 3.5 and 3.6.

3.5.2 Refined anatomically informed midsaggital model of the vocal tract



(a) Geometrical model.

(b) Intensity image corresponding to the geometrical model.

Figure 3.10: Upper airway object model.

In this section we will propose a method to mitigate the potential danger of the contour detection optimization algorithm to get stuck in a local minimum while addressing the specifics of the midsaggital view of the human vocal tract. As outlined in Section 3.4 our model for the midsagittal upper airway image consists of three homogeneous regions on a square background. We now divide the boundary polygon of each region into sections that correspond to anatomical entities as shown in Figure 3.10(a)[5]. This implies that the vertices of a particular boundary section are likely to move in concord during the

[5]The number of polyline vertices in each section was chosen manually so as to capture the underlying geometry with reasonable detail. While [18, 52] showed that a minimum description length criterion can been used to determine the optimum order of B-spline and Fourier contour descriptors, at this point, we leave this issue outside the scope of this thesis.

deformation process. Additionally, we may hypothesize, based on the anatomy, that for a given section particular types of deformation are more prevalent than others. We distinguish the "rigid" deformations of translation, rotation, and scaling from the "non-rigid" complementary deformations. We will explain in Section 3.5.3 how we exploit this knowledge about a section's deformation space with our hierarchical optimization algorithm by selectively boosting and/or inhibiting the gradient descent with respect to the individual deformation components.

As an example, one may consider the boundary section denoted P_{11} in Figure 3.10(a), which corresponds to the hard palate. Since the hard palate is a bony structure covered by a thin layer of tissue it does not change its shape during speech production data collection. However, a possible in-plane head motion of the subject can change its position in the scene through translation, rotation or a combination of both.[6]

The other mostly translating and rotating sections are the mandible bone cavity (P_3), the chin (P_5), the front of the trachea (P_6), the pharyngeal wall (P_7), the upper, right-hand and lower edge of the chosen field of view (P_8, P_9, P_{10}), the bone structure around the nasal cavity (P_{13}), and the nose (P_{14}). While the lips (P_4, P_{15}) may undergo strong non-rigid deformations, the cross-sectional area visible in the midsagittal plane is assumed to be constant, i.e., in general they will not be subject to a strong scaling deformation.

The epiglottis (P_1) and the velum (P_{12}) are special in that they are articulated structures, i.e., they are flexible tissues that are attached to the surrounding structure at one end, where the attachment point is the rough center of rotation. This is illustrated in Figure 3.11, and the mathematical consequences for the model are explained in Section 3.5.3. However, neither the epiglottis nor the velum is expected to scale during the vocal tract shaping process since the amount of tissue in the scan plane is largely constant.

[6]Of course a head motion by the subject perpendicular to the scan plane can very well change the observed shape of the palate, but the subject's head is properly immobilized through secure sideways padding inside the MR receiver coil.

Lastly, the tongue (P_2) undergoes the most complex deformation of all components. Translation and rotation can be caused by changing the mouth opening, but scaling must also be considered since the tongue tissue can move laterally into and out of the scan plane [55]. And given the complex muscular structure of this organ, we expect large deformations complementary to translation, rotation, and scaling as well.

Table 3.3: Region R_1 boundary sections, level 3 boosting factors

Section	Feature	λ_T	λ_R	λ_S
P_1	epiglottis	1	10	0
P_2	tongue	1	1	1
P_3	lower teeth / mandible	1	1	0
P_4	lower lip	1	1	0
P_5	chin	1	1	0
P_6	front of trachea	1	1	0

Table 3.4: Region R_2 boundary sections, level 3 boosting factors

Section	Feature	λ_T	λ_R	λ_S
P_7	pharyngeal wall	1	1	0
P_8	FOV upper boundary	1	1	0
P_9	FOV right boundary	1	1	0
P_{10}	FOV lower boundary	1	1	0

Table 3.5: Region R_3 boundary sections, level 3 boosting factors

Section	Feature	λ_T	λ_R	λ_S
P_{11}	hard palate	1	1	0
P_{12}	velum	0	10	0
P_{13}	nasal cavity	1	1	0
P_{14}	nose	1	1	0
P_{15}	upper lip	1	1	0

All individual sections and their deformation boosting factors corresponding to the specific anatomical characteristics are listed in Tables 3.3, 3.4, and 3.5. The boosting factors are λ_T, λ_R, and λ_S for translation, rotation, and scaling, respectively, and the next section will describe how they are used. In general, the factors are unity if the

corresponding deformation type is expected for the given section, and zero otherwise. Exceptions are the epiglottis and the velum, which are articulated structures. They receive a boosted rotational component, with $\lambda_R = 10$. This choice was experimentally found to speed up the convergence of the optimization algorithm without compromising the stability.

3.5.3 Hierarchical gradient descent procedure

Based on our anatomical model we can now use a hierarchical optimization procedure to improve the fit of the model's geometry to the observed image data through a gradient descent. As outlined above, the main challenge here is to ensure that the optimization flow does not get trapped in a local minimum of the objective function. The authors of [14, 16] demonstrate that in the realm of irregular mesh processing a practical solution to this type of problem is the design of optimizing flows with "induced spatial coherence," and we utilize the same concept in our application.

Upon an initialization with a manually traced subject-specific vocal tract geometry representing a fairly neutral vowel posture, e.g., roughly corresponding to the [ɛ] vowel, we optimize the fit of our model's geometry independently for each MR image. The optimization procedure according to Equation 3.16 is carried out in four separate consecutive stages, each of which features a particular modification of the gradient descent flow:

- Level 1 - allowing only translation and rotation of the *entire 3-region model* geometry, thereby compensating for the subject's in-plane head motion,

- Level 2 - allowing only translation and rotation of *each region's boundary*, thereby fitting the model to the rough current vocal tract posture,

- Level 3 - allowing only rigid transformations, i.e., translation, rotation, and scaling, of each *anatomical section* obeying the section-specific boosting factors of Tables 3.3, 3.4, and 3.5,

- Level 4 - independent movement of all *individual vertices* of all regions.

Thus this approach tries to find a good global match first and then zooms into optimizing smaller details.

In order to describe the mathematical procedure for the boosting of the rigid deformations of a particular contour polygon it is first necessary to define a point which serves as the center for the rotation and the radial scaling. With the exception of the articulated structures in our model, i.e., the epiglottis and the velum, we choose the center of mass of the polygon of interest. Denote with $\mathbf{v}_{i_j} = [x_{i_j}, y_{i_j}]$ the vertex vector j of the boundary polygon P_i. The center or mass of P_i is

$$\bar{\mathbf{v}}_i = \frac{\sum_{\forall \mathbf{v}_{i_j} \in P_i} l_{i_j} \mathbf{v}_{i_j}}{\sum_{\forall \mathbf{v}_{i_j} \in P_i} l_{i_j}} \tag{3.20}$$

where

$$l_{i_j} = \frac{\|\mathbf{v}_{i_{j+1}} - \mathbf{v}_{i_j}\| + \|\mathbf{v}_{i_j} - \mathbf{v}_{i_{j-1}}\|}{2} \tag{3.21}$$

is the length of the polyline section associated with vertex $\mathbf{v}_{i_j}$. Denoting with $\mathbf{F}_{i_j} = \left[\frac{\partial J}{\partial x_{i_j}}, \frac{\partial J}{\partial y_{i_j}}\right]$ the force on a vertex as derived in Equation 3.17, we find the net translation force on a polygon section P_i

$$\bar{\mathbf{F}}_i = \sum_{\forall \mathbf{v}_{i_j} \in P_i} \mathbf{F}_{i_j} \tag{3.22}$$

and similarly, the net rotation torque on a polygon can be computed as

$$\bar{\mathbf{T}}_i = \sum_{\forall \mathbf{v}_{i_j} \in P_i} \mathbf{F}_{i_j} \times (\mathbf{v}_{i_j} - \bar{\mathbf{v}}_i) \tag{3.23}$$

and it is clear that it has only a z-component denoted $\bar{T}_i$. Lastly, the net radial scaling force on a polygon is

$$\bar{K}_i = \sum_{\forall \mathbf{v}_{i_j} \in P_i} \mathbf{F}_{i_j} \cdot (\mathbf{v}_{i_j} - \bar{\mathbf{v}}_i) \tag{3.24}$$

With these formulas we are now in a position to replace the direct gradient descent of Equation 3.19 by a transformed gradient descent step

$$\begin{aligned} \mathbf{v}_{i_j}^{(n+1)} = \left(1 - \epsilon\lambda_S \bar{K}_i^{(n)}\right) \cdot & \\ \begin{bmatrix} \cos\left(\epsilon\lambda_R \bar{T}_i^{(n)}\right) & \sin\left(\epsilon\lambda_R \bar{T}_i^{(n)}\right) \\ -\sin\left(\epsilon\lambda_R \bar{T}_i^{(n)}\right) & \cos\left(\epsilon\lambda_R \bar{T}_i^{(n)}\right) \end{bmatrix} & \left(\mathbf{v}_{i_j}^{(n)} - \bar{\mathbf{v}}_i^{(n)}\right)^T \\ & - \epsilon\lambda_T \bar{\mathbf{F}}_i^{(n)} + \bar{\mathbf{v}}_i^{(n)} \end{aligned} \tag{3.25}$$

Here we rotated the vertex position first, then scaled it, and finally added a translational displacement. Each step can be controlled with the section-specific boosting factors λ_T, λ_R, and λ_S for translation, rotation, and scaling, respectively. The boosting factors, the step width ϵ, and the partitioning of the boundaries of our 3-region vocal tract model into connected sections are dependent on the current hierarchical optimization level, details of which are elaborated below.

Level 1

For the top level we set $\lambda_T = 1$, $\lambda_R = 1$, and $\lambda_S = 0$, and all vertices of $R_{1\ldots3}$ are considered as belonging to a single polygon. The center of rotation is computed with Equations 3.20 and 3.21. An initial stepwidth of $\epsilon = 0.0005$ was chosen, and a simple variable stepwidth algorithm was used in order to achieve faster convergence and avoid oscillations near the minimum. If an iteration step did not lead to a decrease of the objective function J, then its result is discarded, the stepwidth ϵ is cut in half, and the

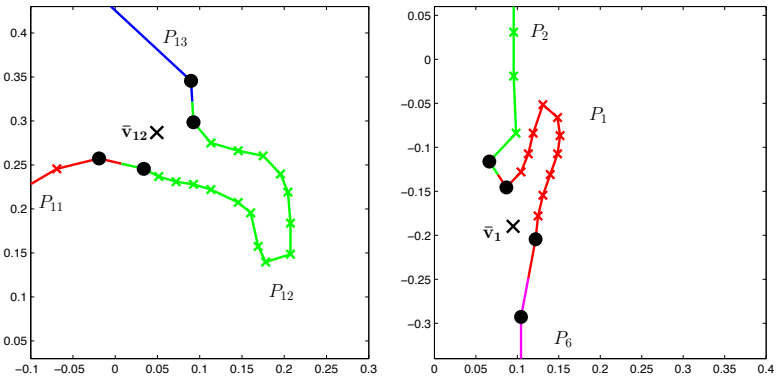
iteration step is carried out anew. In our experiments, a total of 10 iteration attempts are made, and the algorithm then proceeds with level 2.

Level 2

For this stage of the optimization the same transformed gradient descent method is used except that each individual region's complete boundary is now considered a single polyline segment. Here ϵ was set to 0.001, $\lambda_T = 1$, $\lambda_R = 1$, and $\lambda_S = 0$ so that the gradient descent is limited to independent translation and rotation of the regions $R_{1...3}$. The center of rotation is again computed with Equations 3.20 and 3.21. A total of 40 gradient descent steps are attempted.

Level 3

We now break the 3 regions' boundaries into the segments $P_{1...15}$ corresponding to the anatomical components introduced in section 3.5.2, and we apply the boosting factors that are listed in Tables 3.3, 3.4, and 3.5 for the gradient descent with Equation 3.25. Since



(a) Velum. (b) Epiglottis.

Figure 3.11: Velum and epiglottis are articulated structures.

we consider the velum and epiglottis as articulated structures, their centers of rotation $\bar{\mathbf{v}}_{12}$, and $\bar{\mathbf{v}}_1$, respectively, are found with the help of the neighboring (non-articulated)

structures as shown in Figure 3.11(a) and Figure 3.11(b) as black x-markers. The centers of rotation are computed as the mean of the connection vertices (black dot markers). For all other segments we determine the center of rotation and scaling using Equations 3.20 and 3.21. We set $\epsilon = 0.001$ and allow 300 gradient descent attempts.

Level 4

Lastly, we carry out a direct gradient descent according to Equation 3.19 with a step-size $\epsilon = 0.05$, and we allow a total of 300 iteration steps. At this point also non-rigid deformations are accommodated and no manipulation of the gradient descent is done.

3.5.4 Implementation of higher-level contour constraints

When applying the contour detection process described so far no constraints on the boundary polygons are implemented and the vertices move freely during the last optimization stage. At this point we identify two potential problems that can appear during the gradient descent optimization. These problems are possible region overlap, and boundary polygon self-intersection. In general, one can address such issues on a lower level by adding appropriate penalty terms to the objective function, or one can combat these problems at a higher level by directly checking and correcting the geometrical model after each gradient descent step. Our general approach is to operate at a higher level to circumvent the potentially difficult evaluation of such penalty terms and their respective gradients.

The three tissue areas in our upper airway image model are naturally always non-overlapping regions but occlusions happen frequently in the vocal tract during speech production. In these cases two adjacent tissue regions of the model appear connected and the unconstrained gradient descent procedure may in fact let the regions grow into one

another. We handle this issue by detecting after each gradient descent step if a vertex has been moved into a neighboring region's boundary polygon, and if so it is reset in the middle of the two closest correct neighboring boundary vertices.

At this point our algorithm does not include a way of handling self-intersections of the boundary polygons but in practice this seems not to be an issue as the provided sample images will show. Here the reader should keep in mind that our final goal is to be able to extract the vocal tract aperture, and as long as possible self-intersections in the boundaries do not hinder this process we will tolerate them.

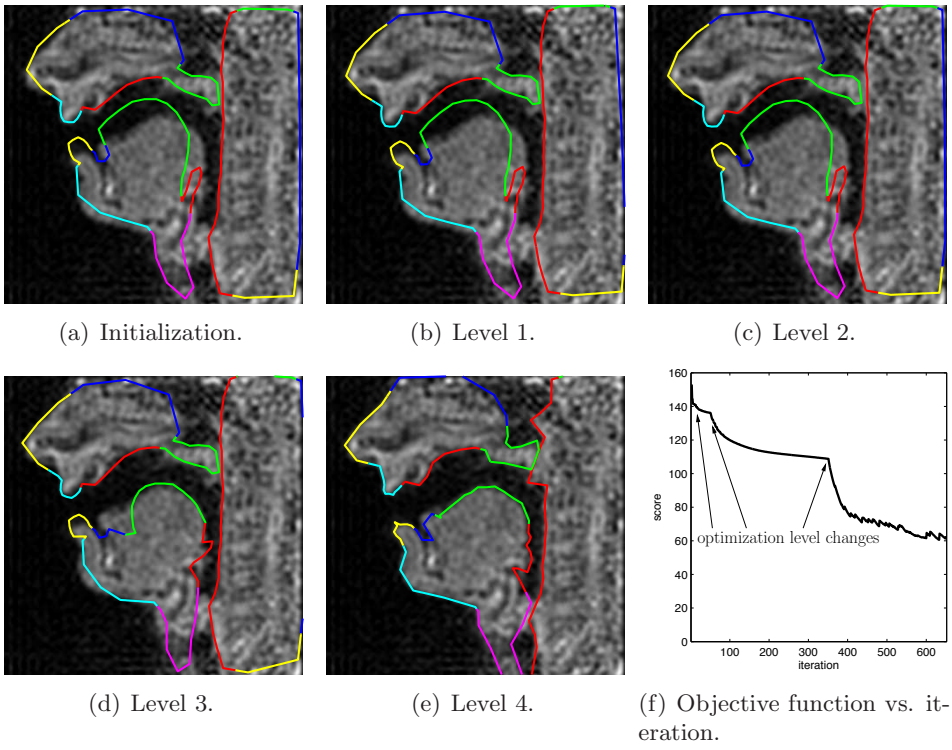
3.5.5 Validation of the hierarchical contour detection algorithm

In this section we present simulation results to demonstrate the effectiveness of our method, and we address the verification of the algorithm. Here we face three problems when trying to assess the achievable accuracy with the proposed algorithm. First, we have no information on the true underlying contours of a particular image. A comprehensive upper airway MR phantom does not exist, and the validity of a manually traced reference of real-time MR images is doubtful given the poor quality in terms of image noise. We have already shown in Section 3.4.3 that manual tracings can exhibit larger variations and are less accurate than automatic tracings. Second, we would need a measure that captures the difference between a reference boundary and the boundary found by the algorithm, as well as the accuracy with which the individual anatomical sections of the boundary contours are found. Third, as pointed out in Section 3.5.2, we currently have fixed the number of vertices for the contour representation, and the achieved fit of the model to the observed image data is dependent on the number of degrees of freedom.

At this point, we present a variety of sample images that have been processed with our algorithm, and we will qualitatively judge the detected contours and analyze the

results from an intended end-use application point of view. All images presented in this paper were extracted from a recording of German read speech, and we refer the reader to download[7] and view the entire video `movie1_processed.mov`.

Lastly, we have made available online two other MR video sequences `movie2.mov` and `movie3.mov` and their corresponding contour-traced versions `movie2_processed.mov` and `movie3_processed.mov`, which contain English read speech from two female subjects. These two sequences were recorded with an appropriate zoom factor setting since the female vocal tract is in general smaller than the male vocal tract. All segmentations were automatically derived using the algorithm presented in this paper. The audio track for all movies was obtained concurrently using the procedure described in [9].



(a) Initialization. (b) Level 1. (c) Level 2.

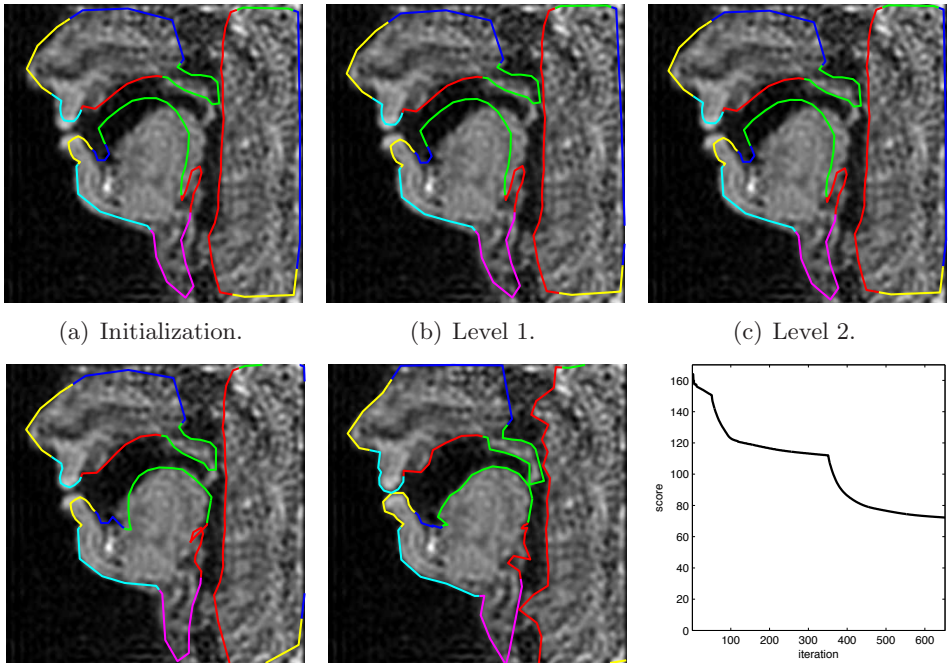
(d) Level 3. (e) Level 4. (f) Objective function vs. iteration.

Figure 3.12: Example 1: Vowel [ɑ] extracted from a read speech sequence

[7] `http://sail.usc.edu/span/tmi2008/index.php`

For the purpose of evaluation we consider four different images that contain distinct vocal tract postures obtained from German speech in a real-time MR scan. The posture shown in Figure 3.12 belongs to the vowel [ɑ]. Figure 3.12(a) shows the image and the initialization contour, whereas Figure 3.12(b), Figure 3.12(c), Figure 3.12(d), Figure 3.12(e) show the result at the end of stages 1, 2, 3, and 4 of the optimization algorithm, respectively. We can see that at the end of the last stage, the outline of the articulators of interest are well captured. The critical and difficult-to-capture velum posture was detected only partly correct in as the velum falsely appears shortened and the pharyngeal wall appears to bulge. This result is due to the occlusion with the pharyngeal wall, i.e., no air-tissue boundary exists which could be found by any edge detection algorithm. Putting a constraint on either the rigidity of the pharyngeal wall contour, or the constant cross-sectional area of the velum could improve this result. However, the velum and pharyngeal wall contours do touch and the occlusion was recovered correctly, and hence the zero velum aperture would be detected properly. Figure 3.12(f) shows the evolution of the objective function J over the course of the 650 iterations. The major discontinuities in the curve correspond to the changing of the iteration level, whereas the noise-like perturbations during the final 300 iterations are caused by the higher-level contour clean-up procedure which keeps removing overlap between velum and pharyngeal wall.

Figure 3.13 is an example with a bilabial closure corresponding to the nasal [m]. Notice that the lowered velum posture was identified correctly and the corresponding velum aperture tract variable can be estimated meaningfully. Furthermore, we notice that the shape of the glottis was not captured correctly since the spacing of the boundary vertices for the pharyngeal wall was chosen a bit too coarse. As before, the objective function time evolution exhibits distinct jumps at 10, 50, and 350 iterations, which correspond to the changes of the optimization level.



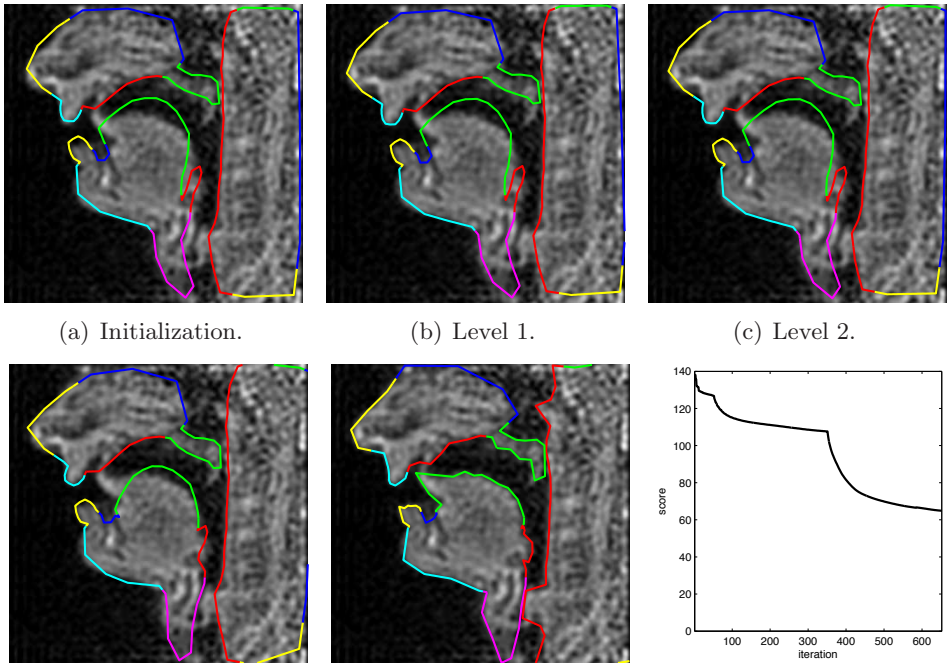
(a) Initialization. (b) Level 1. (c) Level 2.

(d) Level 3. (e) Level 4. (f) Objective function vs. iteration.

Figure 3.13: Example 2: bilabial nasal [m]

Figure 3.14 and Figure 3.15 contain two examples with two different and often occurring types of occlusions in the oral cavity. The former shows the tongue tip touching the alveolar ridge as is the case for the lateral approximant [l], while the second example corresponds to the postalveolar fricative [ʃ]. In both cases the critical tongue posture was found to be consistent with what is expected for the articulation of these speech sounds.

The 25 images corresponding to the spoken sequence [lasɛn] are shown in Figure 3.16. We can see that the lips, the velum, and the tongue position and shape are captured well in spite of the poor quality of the image data.



(a) Initialization. (b) Level 1. (c) Level 2.

(d) Level 3. (e) Level 4. (f) Objective function vs. iteration.

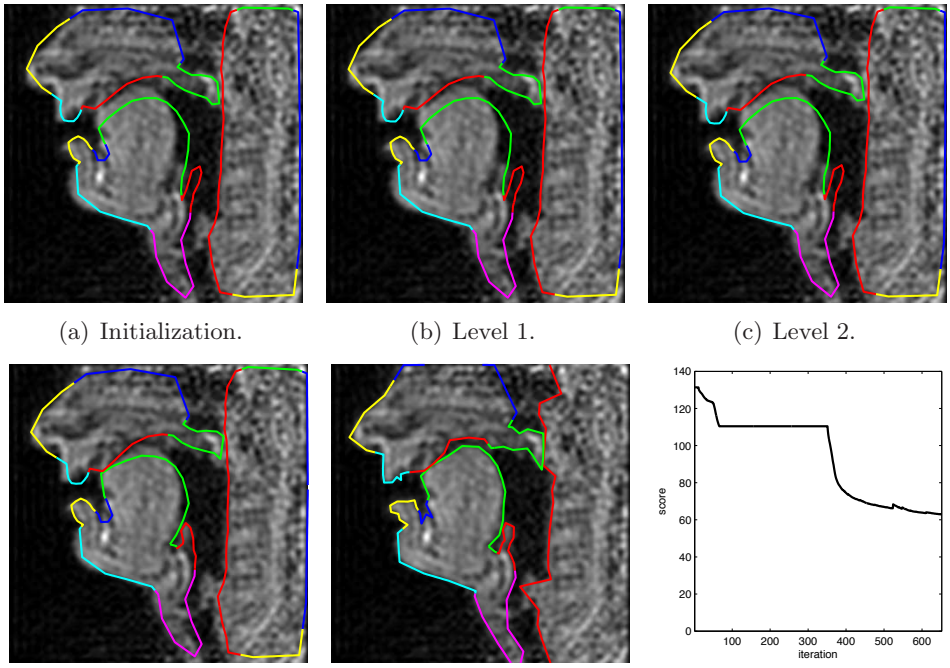
Figure 3.14: Example 3: lateral approximant [1]

3.6 Discussion

3.6.1 Summary

In this chapter we presented a spatial frequency domain-based method for unsupervised multi-region image segmentation with application to contour detection. The formulation in the spatial frequency domain allows the direct analytical closed-form computation of the external energy of the boundary contour as well as its gradient with respect to the contour parameters. The key mathematical ingredient to this approach is the closed-form solution of the 2-dimensional Fourier transform of polygonal shape functions, and it affords the continuous valued optimization of the boundary contour parameters.

The spatial frequency domain image data can be obtained through MR acquisitions directly, or through the Fourier transform of any general pixelized image data stemming



(a) Initialization. (b) Level 1. (c) Level 2.

(d) Level 3. (e) Level 4. (f) Objective function vs. iteration.

Figure 3.15: Example 4: postalveolar fricative [ʃ]

from any other image sampling method, e.g., digital photography or raster scanning. The operation in the Fourier domain circumvents any fine-grain interpolations of such raster-sampled image data, and it is equivalent to operating on an ideally sinc-interpolated version of the rasterized image.

For the case of medical image processing, our segmentation algorithm introduces a general framework for incorporating an anatomically informed object model into the contour finding process. The algorithm relies on the manipulation of the gradient descent flow for the optimization of an overdetermined non-linear least squares problem.

We quantitatively evaluated our method using direct processing of MR data obtained from a phantom experiment with a relatively simple geometry. Our method was found to outperform careful manual processing in terms of achieved accuracy and variation of the results. We furthermore demonstrated the effectiveness of our algorithm for the case

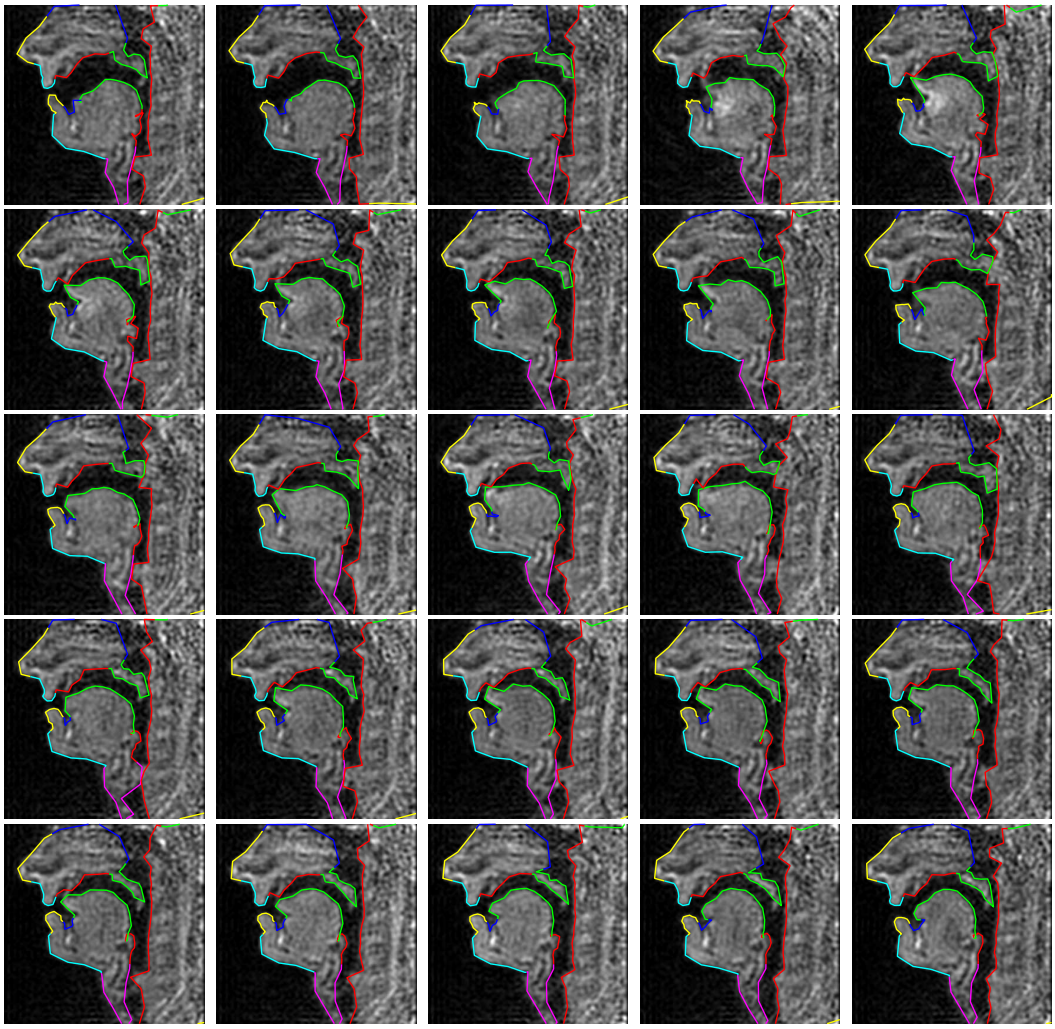


Figure 3.16: Sequence [lasɛn] in 25 images (from left to right, top to bottom).

of air-tissue boundary detection in midsagittal real-time MR images of the human vocal tract, and we illustrated the method using a variety of sample images representing vocal tract postures of distinct phonetic quality.

The algorithm is applied to individual images, and it is hence useful for parallel processing of image sequences on computing clusters. After determining a general upper airway initialization contour for a subject, which takes about 5 minutes for the selection and manual tracing of a suitable frame, the algorithm is unsupervised. For each upper

airway image the processing time for 650 gradient descent optimization steps was approximately 20 minutes using a MATLAB[8] implementation on a modern desktop computer. If longer processing time is tolerable better results can be achieved by using more iterations, smaller step width, and more densely spaced vertices to allow to capture even finer details of the boundary contours.

3.6.2 Open research questions

Our framework leads to a variety of new open research questions, one being how to quantitatively assess the achievable segmentation accuracy for in vivo upper airway data. This requires constructing a realistically moving upper-airway MR phantom which is capable of mimicking the tissue deformations typical for speech production.

Also of interest are the development of procedures to combine directly the k-space data of multiple MR receiver coils so as to circumvent the current pre-processing step in the discretized image domain.

Furthermore, it is an open question if there are any other boundary contour descriptors which have a closed form mathematical solution to the 2-dimensional Fourier transform of the corresponding shape function, and ideally are also inherently self-intersection free. While it is possible to approximate smooth boundaries with the help of densely spaced polyline vertices, and detect and remove self-intersections using a higher level process there is a price in terms of computation time.

Another research avenue could lead to a more sophisticated gradient descent procedure for solving the inherent optimization problem, possibly using implicit Euler integration. Since we have the capability for polygonal boundaries to analytically evaluate the Hessian

[8] `http://www.mathworks.com/`

of the objective function along with the gradient the application of the Newton-algorithm may be a way to speed up the algorithm's performance as well.

Yet another improvement to the method could possibly be made by imposing more refined constraints on the boundary contours, such as roughness penalties, or a constant enclosed area constraint. An example would be the velum, whose midsaggital cross-sectional area can be considered constant. For the tongue, on the other hand, such a constant area constraint would probably not work since its volume can move in and out of the midsaggital scan plane. Such additional constraints can be implemented on a low level by adding internal energy terms to the objective function. It is also possible to constrain at a higher level the geometrical object model directly.

Furthermore of interest are the extensions to other related imaging applications such as coronal or axial vocal tract real-time MR images, as well as models for other medical applications, e.g., cardiac real-time MR imaging. Corresponding to the available image data, new object models could be either 2- or 3-dimensional. Notice that the necessary closed-form equation for the 3-dimensional Fourier transform has been derived in [41] as well. This also makes possible the construction of geometrical models that do not only include constant amplitude regions but even areas with spatially linearly varying intensity.

Moreover, future research could be directed towards the application of our framework for MR acceleration using the methods of compressed sensing MRI. There the objective will be to reduce the over determinedness of the optimization problem and utilize fewer MR data samples to accomplish the region segmentation task, which can lead to a shortened acquisition time and/or higher frame rate.

Chapter 4

RT-MRI investigation of resonance tuning in soprano singing

4.1 Abstract

This section investigates using RT-MRI the vocal tract shaping of 5 soprano singers during the production of two-octave scales of sung vowels. A systematic shift of the first vocal tract resonance frequency with respect to the fundamental is shown to exist for high vowels across all subjects. No consistent systematic effect on the vocal tract resonance could be shown across all of the subjects for other vowels or for the second vocal tract resonance.

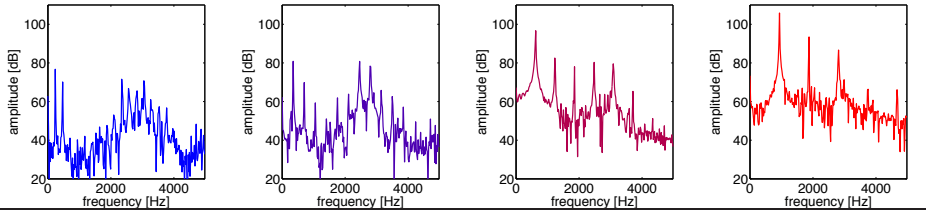
4.2 Background

The singing voice has been of considerable interest to the acoustics researcher for a long time, and in particular the concept of resonance tuning has drawn notable attention over the past decades [13, 63]. Resonance tuning is a strategy that trained opera singers are hypothesized to employ in order to increase their vocal efficiency and output power. Before the availability of audio power amplification this was an obvious necessity when performing in large concert halls.

During a vocal song production, the artist faces at least three constraints. Besides the need for an adequate intensity, the pitch at any given point in time is dictated by the melodic score of the music. Furthermore, the lyrics of the song have to be rendered with some degree of fidelity, which in turn demands the maintenance of the linguistic identities of the sung sounds (e.g., vowels) to some extent [58].

The theory of resonance tuning now contends that the vowel identity requirement is relaxed in practice and that trained singers actively modify their vocal tract shape so as to shift one of the resulting resonance frequencies to a multiple of the current (target) pitch frequency [60]. So, even though the changed formant structure alters the vowel quality, the singer is able to maintain the pitch in accordance with the score of the music while simultaneously maximizing the voice output.

Table 4.1: 1024-point FFT spectra for /i/ at notes 1, 5, 11, and 15 (subject M1).

F_0 [Hz]	233	349	622	932
spectrum				

Showing evidence for resonance tuning using audio recordings alone is not straightforward since the estimation of vocal tract resonance frequencies can be difficult, in particular for the case of high-pitched singing, e.g., soprano singing [27]. Here, the glottal source spectrum contains much wider spaced harmonics than in normal speech, so that the estimation of the resonance frequencies from peaks in the spectral envelope of the recorded signal is severely compromised (see, for example, Table 4.1). Therefore, researchers have resorted to other methods for the investigation of the vocal tract transfer function.

One possibility is the use of an artificial external broad-band noise source to excite the vocal tract while the soprano singer tries to maintain her natural singing vocal tract posture without actually producing any sound [26]. Subsequently, a resonance frequency estimation can be carried out from the reflected sound waves.

Another option is to obtain direct evidence of the vocal tract shaping strategies auch as using MRI [61, 59]. However, to acquire a conventional (static) MRI recording the singer may have to hold the vocal tract posture for an unusually long time, e.g., on the order of a few minutes as would be the case for a high resolution 3-D volumetric scan. To alleviate this issue researchers often restrict themselves to capturing the midsagittal view of the vocal tract and then performing an aperture-to-area function conversion to facilitate a tube model description of the vocal tract. However, even a 2-D static MRI scan can easily take a few seconds.

In contrast to the previous studies, this study employs RT-MRI technology to obtain midsagittal vocal tract image data from a total of 5 soprano singers. While thus far RT-MRI has been mostly used to study dynamic speech production processes, it also appears well suited for the investigation of scale singing since it allows the subjects to produce vocal sounds in a more natural way, i.e., they are not required to maintain the vocal tract posture for unnaturally long periods of time [6].

Furthermore, RT-MRI allows the researcher to investigate other aspects of song productions, such as their expressive qualities, rhythm and pausing behavior, etc., which require data from dynamic productions. Though this article focuses on sung vowel scales, it does describe the data acquisition, processing, and analysis steps relevant for general song production (data examples can be found in[1]). In that regards, it can be viewed

[1] http://sail.usc.edu/span/

as providing a proof-of-concept for the use of RT-MRI technology for studies of vocal productions of song.

4.3 Data collection

The subjects for this study were 5 female sopranos (M1, S2, K3, L4, H5) trained in Western opera and who were native American English speakers. The subjects sang two-octave vowel scales (/la/, /le/, /li/, /lo/, /lu/) without vibrato, and they were allowed to breathe after the first octave.

Midsagittal MR images were collected with a GE Signa 1.5T scanner [45]. Synchronized audio recordings were obtained, and the scan noise was subsequently removed [9]. During the data collection the subjects were in a supine position.

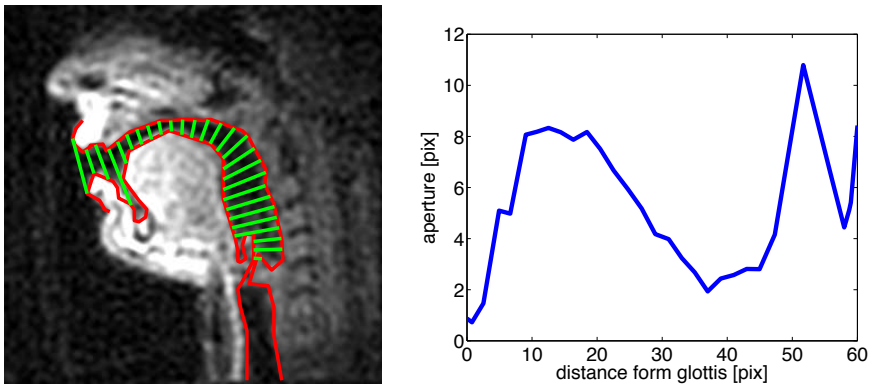
4.4 Data analysis

4.4.1 Audio analysis

Using the noise-cancelled audio recording, a pitch estimation was carried out using the PRAAT software[2]. However, as described above, the estimation of the vocal tract resonances from the audio signal is difficult, especially at high pitch values. This is due to the fact that the harmonics of the source spectrum are widely spaced, and consequently the filter function of the vocal tract gets sampled only at relatively fewer frequency points (see Table 4.1). Therefore, the vocal tract resonance frequencies were estimated directly using the midsagittal image data. And while these estimates can be noisy, we are mainly interested in statistically significant trends of the resonance frequencies with respect to the fundamental.

[2] `http://www.fon.hum.uva.nl/praat/`

4.4.2 Image analysis



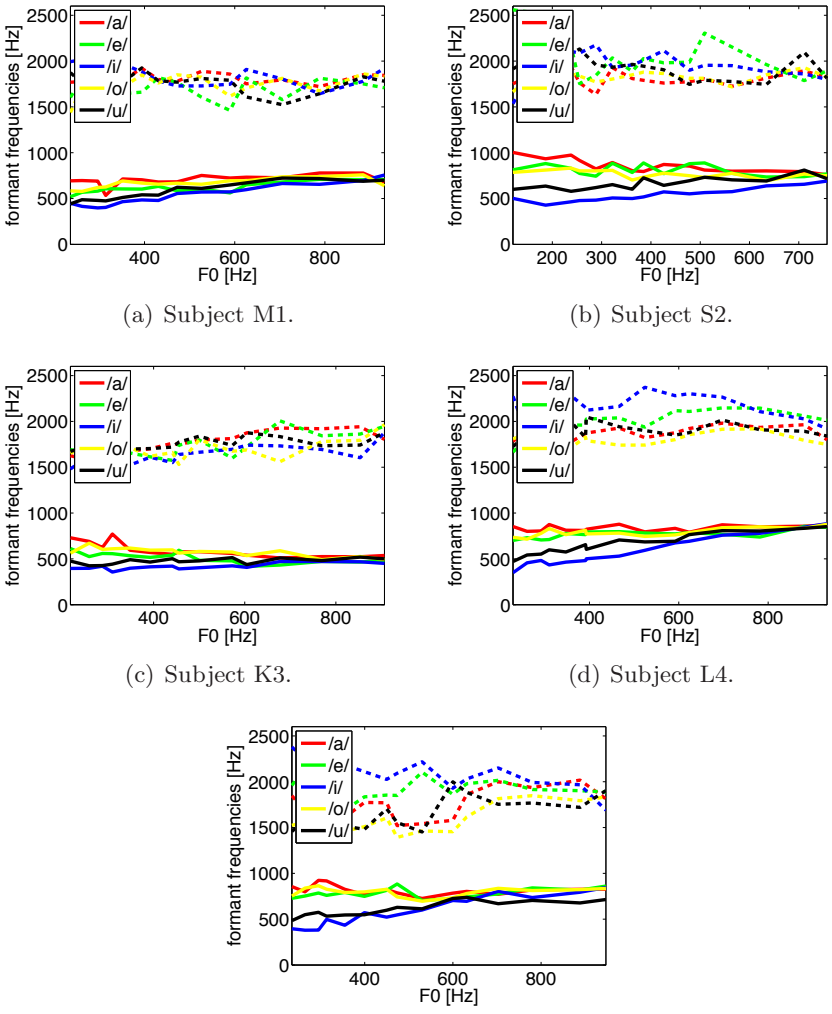
(a) Sample midsagittal real-time MR image.

(b) Aperture function from glottis (left) to lips (right).

Figure 4.1: Subject M1, producing /le/ at note 1.

From each of the notes of the scales, one image was extracted corresponding to the midpoint of the vowel segment, i.e., from a relatively stable vocal tract configuration. In these images the vocal tract outline was then automatically detected[7] and then manually corrected if necessary. The glottis position was manually determined in each image. A sample image is shown in Figure 4.1(a), showing subject M1 singing /le/ at note 1. Here, the vocal tract outline is shown in red.

Subsequently, the aperture function from the glottis to the lips was derived from the vocal tract contours. This was accomplished by first constructing a vocal tract midline using repeated geometrical bisection, and, secondly, finding densely spaced perpendiculars along the midline and their intersections with the vocal tract contours [3]. The perpendiculars are the midsagittal aperture lines, and they are shown in green in Figure 4.1(a). Figure 4.1(b) shows the aperture function corresponding to the vocal tract shape of Figure 4.1(a). This graph displays the length of the aperture lines as a function of position along the midline. In Figure 4.1(b) the left side corresponds to the glottis, while the right side corresponds to the lips. The units used in the graph are pixels.



(a) Subject M1.

(b) Subject S2.

(c) Subject K3.

(d) Subject L4.

(e) Subject H5.

Figure 4.2: Resonances F_1 (solid), and F_2 (dashed) versus the fundamental F_0.

The midsagittal aperture function was then converted to the cross-sectional area function of a tube model whose resonance frequencies were computed using the VTAR[66] software. Figure 4.2 shows the resonances F_1 and F_1 as a function of the fundamental F_0 for all 5 vowels for all 5 subjects. The resonance frequency estimates then form the basis of the statistical analysis in Section 4.5.

It must be pointed out that numerous methods have been proposed for the aperture-to-area conversion and, in general, their optimum parameters are subject specific [54].

For this study the method described in [32] and extended in [35] was employed without adaptation of the parameters. Hence deviations of the computed tube model resonances from the true vocal tract resonances must be expected. However, this study aims at identifying global trends in the formant frequencies with respect to the pitch frequency for a given subject, as opposed to quantifying absolute formant frequency measurements.

4.5 Results

Table 4.2: Sample MR images and midsagittal aperture functions of all 5 vowels at notes 1, 5, 11, and 15 (subject M1).

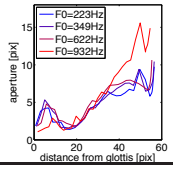
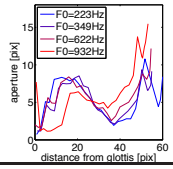
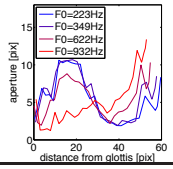
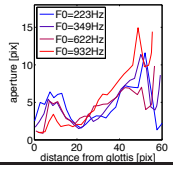
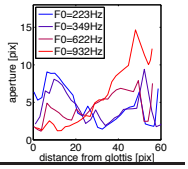
F_0 [Hz]	/a/	/e/	/i/	/o/	/u/
233					
349					
622					
932					
apt. fn.					

Table 4.2 shows the midsagittal images for subject M1 for all 5 vowels at notes 1, 5, 11, and 15 with fundamental frequencies of 233Hz, 349Hz, 622Hz, and 932Hz, respectively. It can be seen that for the low notes the vocal tract configuration is distinct for the individual vowels, and the distinction decreases as the pitch increases. This behaviour was observed for all 5 subjects.

The bottom row in Table 4.2 shows the aperture functions of subject M1 for the 5 vowels for the notes 1 (blue), 5 (dark purple), 11 (light purple), and 15 (red). It can be seen that at higher notes the individual differences between the vowels decrease, and in particular the shape of the oral cavity converges to a widely open configuration.

Corresponding to the /i/-column of Table 4.2, the 1024-point FFT spectra at notes 1, 5, 11, and 15 are shown in Table 4.1, which were derived from the noise-cancelled audio recording. These examples illustrate the difficulty of the estimation of the vocal tract resonances at high pitch values. At the low note 1 resonance peaks can be recognized in the spectrum easily, whereas at the high note 15 no resonances are readily observable.

In order to investigate the dependence of the vocal tract resonances F_1 and F_2 on the fundamental F_0, linear models were fit of the form

$$F_{1,2} = \beta_{1,2} \times F_0 + \alpha_{1,2} + \epsilon \tag{4.1}$$

for each vowel. Here, α has the dimension of [Hz], and β is the dimensionless slope of the regression line. The value ϵ represents the error. The calculated values are listed in Table 4.3, and we also list the resulting p-value for the respective β coefficient. In Table 4.4 we compact this information more, and we list only the sign of the statistically significant trends ($\beta \neq 0$ with significance $\geq 95\%$) for all subjects and all vowels.

These values suggest that for the high vowels /i/ and /u/ for all subjects there is a consistent dependency of the first vocal tract resonance F_1 on the fundamental F_0 in

Table 4.3: Linear regression of the vocal tract resonances versus the fundamental.

Subject	Vowel	F_1			F_2		
		α_1 [Hz]	β_1	p	α_2 [Hz]	β_2	p
M1	/a/	639	0.126	0.061	1769	0.061	0.427
	/e/	506	0.221	3×10^{-5}	1676	0.036	0.783
	/i/	291	0.490	6×10^{-10}	2025	-0.314	0.032
	/o/	580	0.167	0.003	1613	0.230	0.092
	/u/	378	0.405	6×10^{-7}	1884	-0.213	0.088
S2	/a/	975	-0.297	1×10^{-4}	1808	-0.000	0.999
	/e/	851	-0.099	0.21198	2305	-0.598	0.053
	/i/	360	0.425	4×10^{-10}	2133	-0.401	0.115
	/o/	812	-0.108	0.037	1796	0.085	0.412
	/u/	538	0.301	2×10^{-4}	2099	-0.401	0.043
K3	/a/	732	-0.272	6×10^{-4}	1539	0.446	2×10^{-5}
	/e/	603	-0.179	0.003	1516	0.429	0.004
	/i/	357	0.123	4×10^{-4}	1470	0.339	0.002
	/o/	663	-0.178	4×10^{-4}	1582	0.245	0.065
	/u/	431	0.090	0.017	1643	0.217	0.026
L4	/a/	809	0.051	0.186	1782	0.161	0.060
	/e/	692	0.148	0.002	1738	0.465	0.016
	/i/	256	0.671	2×10^{-9}	2269	-0.170	0.352
	/o/	715	0.149	0.002	1784	0.044	0.593
	/u/	418	0.498	2×10^{-8}	1846	0.084	0.464
H5	/a/	853	-0.067	0.282	1579	0.343	0.057
	/e/	729	0.102	0.108	1942	-0.033	0.841
	/i/	237	0.680	2×10^{-8}	2256	-0.393	0.066
	/o/	789	0.016	0.789	1281	0.570	5×10^{-4}
	/u/	460	0.305	5×10^{-5}	1341	0.587	0.002

Table 4.4: Sign of the statistically significant linear trends of the resonances F_1 and F_2 with respect to the fundamental F_0.

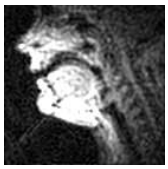
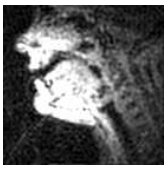
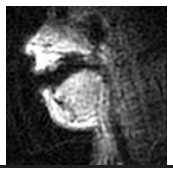
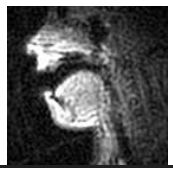
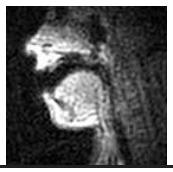
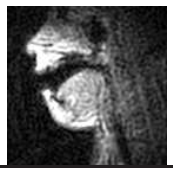
Subject	F_1					F_2				
	/a/	/e/	/i/	/o/	/u/	/a/	/e/	/i/	/o/	/u/
M1		+	+	+	+			−		
S2	−		+	−	+					−
K3	−	−	+	−	+	+	+	+		+
L4		+	+	+	+		+			
H5			+		+				+	+

terms of a positive correlation. Other than that, no clear patterns can be readily observed that apply across all subjects.

4.6 Discussion

The finding that the first resonance of the high vowels rises with the fundamental frequency is consistent with previous findings. Considering the sample images in Table 4.2, it is easy to see that the front cavity opens more widely as the singer goes to higher fundamental frequencies, and it is well known that F_1 is directly related to the opening degree. The relative opening effect is certainly strongest for the high vowels /i/ and /u/, which are most constricted in their natural oral cavity configuration. Hence the quantitative findings are well in accordance with the expectations, and we conclude that the RT-MRI data and the proposed processing steps offer merit.

Table 4.5: MR images for all 5 subjects and all 5 vowels at note 15 (F_0 = 932Hz).

Subject	/a/	/e/	/i/	/o/	/u/
M1					
S2					
K3					
L4					
H5					

However, based on our study, we cannot conclude that all sopranos employ generalizable strategies for resonance tuning the way it has been described in prior literature. To illustrate the qualitative differences in the shaping strategies, we show in Table 4.5 the MR images for all 5 subjects and all 5 vowels corresponding to note 15 (F_0 = 932Hz), which is the highest note in our data set. We observe that in particular subject M1 but also S2 (top 2 rows) show evidence of some of the vowel-specific tongue shaping even at this extreme pitch, whereas the rest of the subjects appear to have converged to a single canonical vocal tract shape for all vowels. Furthermore, the width of the oral cavity varies considerably across subjects, with M1 being on one extreme and K3 on the other.

We speculate that the observed variability in the vocal tract shaping may be due to the individual training that each of the singers had received. In this regard it would be also interesting to see if RT-MRI recordings can be used in the future as a teaching tool for voice teachers to help sopranos acquire consistent tuning strategies. In summary, we find that the interaction between singing and linguistic goals of producing speech sounds is complex and needs further exploration.

Chapter 5

RT-MRI analysis of vocal tract shaping in English sibilant fricatives

5.1 Abstract

This study uses RT-MRI to investigate shaping aspects of two English sibilant fricatives. The purpose of this article is to 1) develop linguistically meaningful quantitative measurements based on vocal tract features that robustly capture the shaping aspects of the two fricatives, and 2) provide qualitative analyses of fricative shaping. Data was recorded in both midsagittal and coronal planes. The proposed three quantitative measures of this study provide robust results in categorizing shape. The qualitative analyses describe tongue shape in terms of grooving and doming and they support previous research.

5.2 Introduction

RT-MRI promises a new means for visualizing and quantifying the spatio-temporal articulatory details of speech production. This study uses RT-MRI to study the shaping and dynamic aspects of two English sibilant fricatives /s/ and /S/. The RT-MRI data, which affords views of the entire moving vocal tract, and is accompanied by synchronized audio recordings, aims to build upon, and add to, the several excellent previous/ongoing

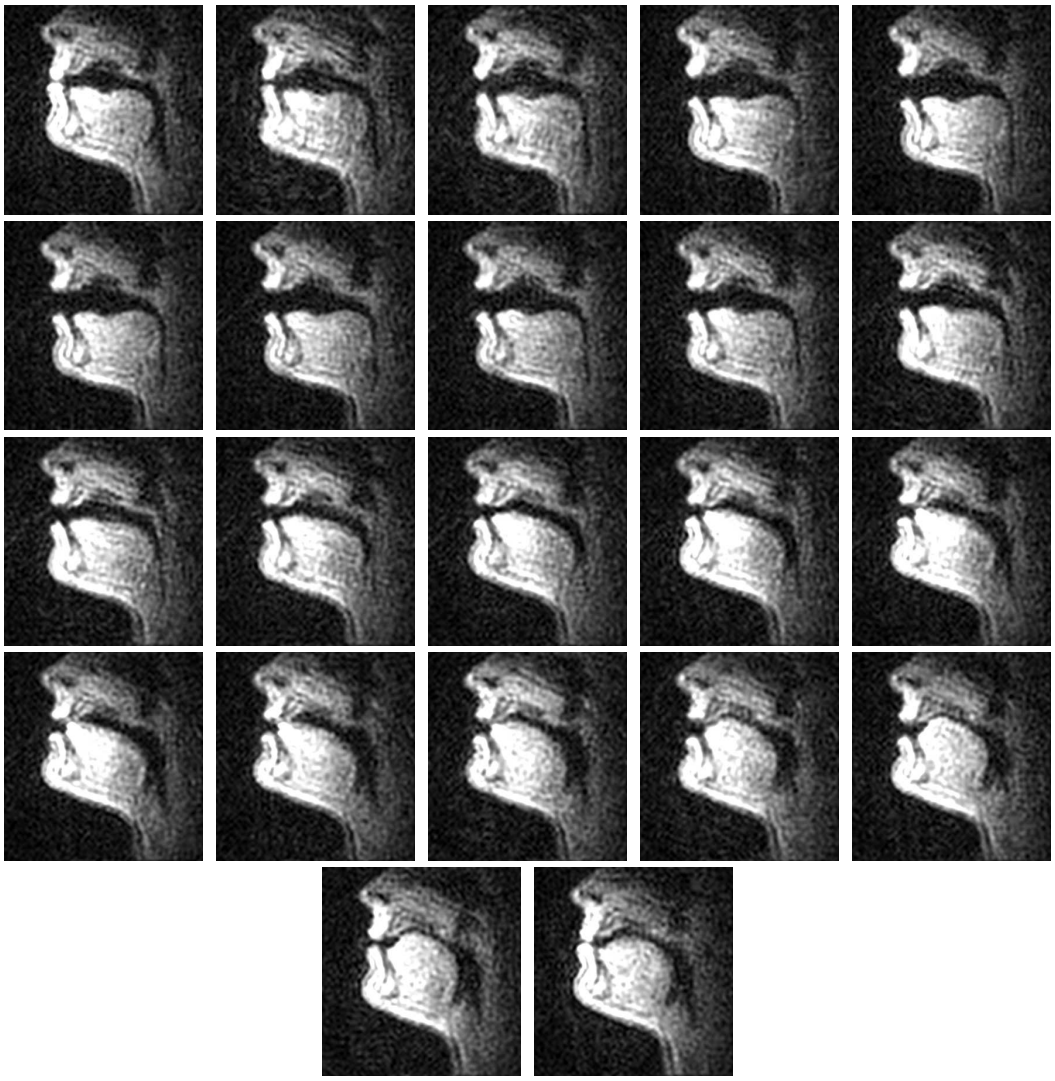


Figure 5.1: Production of "pa seep" by subject S1 in 22 midsagittal images (from left to right, top to bottom).

studies that aim to fully capture the static and dynamic properties of fricatives. Acoustic studies are capable of illuminating the dynamical nature of fricatives, as changes in formant structure indicate changes in the vocal tract while producing speech sounds, (e.g., [1], [53]). However, acoustic studies are unable to characterize the exact shaping of the tongue (and perhaps other articulators) necessary for the production of fricatives. Other experimental methods for obtaining speech production information (e.g. EMA,

static MRI, x-ray, etc.), lack information about change over time, or lack the spatial resolution necessary to properly characterize the complete shape of articulators during the production of fricatives. RT-MRI provides an ideal tool for measuring both the posture and temporal (dynamic) properties of fricatives and the effect of surrounding vowel context on the fricatives, as MRI provides us a (nearly) complete picture of the dynamics of the vocal tract. RT-MRI additionally provides valuable insights into the production of fricatives as it is able to examine the shaping of such fricatives in various planes. The current study employs this technique to obtain data from both midsagittal and coronal planes of the vocal tract.

Specifically, the aim of this paper two-fold. The primary objective is to investigate various derived measurements from MR images that are linguistically meaningful. These measurements are based on tongue shape and other properties of the vocal tract seen in the MR images, and they allow for analysis of the two fricatives under examination. The measurements described below allow for the explicit study of shaping differences between the tongue tip and tongue body gestures of /s/ and /S/. Sibilant fricatives are of particular interest because of the complexities displayed in their shape during production. This shaping of fricatives (e.g. the grooving of the tongue for sibilant fricatives) has shown to be crucial in yielding their acoustic properties necessary for perception [37]. Thus, defining measurements that allow for shape to be investigated as a variable is crucial for the validation of any hypothesis addressing the role of shape in the production of fricatives. Three derived measures are shown to yield linguistically meaningful results. They are 1) tongue-palate area behind fricative constriction (midsagittal), 2) tongue-palate area deformation (midsagittal), and 3) fricative groove-depth (coronal). These measurements are discussed in the methods section.

Secondly, the study aims to describe various strategies used by speakers to produce sibilant fricatives. Based on previous research on the shaping of English fricatives (e.g., [43]) it is expected that /s/ will show grooving of the tongue and /S/ will show doming of the tongue. This study also asks what effect surrounding vowel context will have on the shaping of the fricatives. The fricatives under examination were preceded and followed by the vowels /i/ and /a/. The vowel /i/ has been described as showing a convex dome-like shape, whereas /a/ is considered to be flat [44]. The study thus seeks to determine what happens when two contrasting shapes (such as flat /a/ and grooved /s/) are adjacent. Two possible hypotheses are considered: target shaping in the fricative will be more pronounced when contrasting shapes are adjacent, or target shaping in the fricative will be more pronounced when comparable shapes are present.

Several previous studies using MRI have shown shape as crucial parameter in the production of fricatives. Co-articulatory effects have been shown to be important in describing the shape of Swedish fricatives [17]. When examining English sibilant fricatives with MRI, various observations about tongue posture have been shown to be important. /S/ has been shown to be articulated with a raised tongue blade that is distributed across the alveolar ridge. /S/ also shows a sublingual cavity behind the constriction, whereas [s] shows an absence of a sublingual cavity [49]. The concave shaping of the tongue has also been shown to be crucial for the production of English sibilant fricatives; /s/ consistently shows concavity behind the constriction region, whereas /S/ does not [43]. The area posterior to the constriction has also been shown to be important: the area functions derived for /s/ tend to be less smooth than those for /S/. This difference has been shown to be due to a slight raising of the tongue for /s/ postures. This study further investigates the importance of the area behind the constriction during fricatives, examining area and

area deformation behind the tongue. These two variables are shown to be important to quantitatively characterize the shaping differences in the two fricatives.

5.3 Methods

Three native speakers of American English were used as subjects, two female (A2, S2), and one male (A1). Subjects had no known speech or hearing deficits.

5.3.1 Stimuli

The fricatives /s/ and /S/ under analysis occur between either the vowels /i/ or /a/ in a carrier phrase ("Go pVCVp okay", V∈{/a, i/}, C∈{/s, S/}). There are four different surrounding vowel contexts: symmetrical: /i_i/, /a_a/; asymmetrical: /i_a/, /a_i/. As there are two different fricatives, this yields eight different stimuli.

The sentences were grouped into fourteen blocks, seven blocks for each fricative. Thus each stimulus was read seven times by each subject. Each block containing four sentences was randomized, and the order of the blocks was randomized, with the constraint that blocks alternated according to fricative (two /s/ blocks were never consecutive).

5.3.2 RT-MRI and synchronized audio acquisition

MR images were acquired on a GE Signa 1.5 Tesla scanner using a fast gradient echo pulse sequence with a 13-interleaf spiral readout [45] within the RTHawk framework [51]. A four-channel targeted phased-array receiver coil was employed. Images were formed from the data of two coils located in front of the subject's face and neck through root sum of squares combining. The repetition time TR was 6.376ms. The MRI reconstruction was carried out using a standard gridding and sliding-window technique [25] with a window

offset of 7 acquisitions. The resulting in a frame rate for processing and analysis was 22 frames per second. The slice thickness was 3mm.

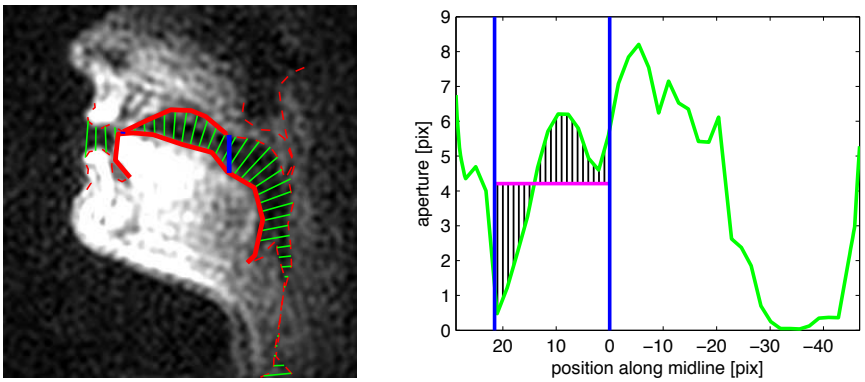
Images were acquired in the midsagittal plane and in a coronal plane. The midsagittal field-of-view (FOV) was chosen to capture the entire vocal tract from glottis to lips. The image rotation was chosen so that the pharyngeal wall is approximately vertical. A sample midsagittal image sequence of the utterance "pa seep" is shown in Figure 5.1. The coronal scan plane was selected perpendicular to the midsagittal scan plane at the position of maximal doming of the hard palate. The FOV was identical in size with respect to the midsagittal value. A sample coronal image is shown in Figure 5.4.

Simultaneous synchronized speech audio was collected during the MRI scans. Subsequently, a noise cancellation procedure was applied to the audio signal to remove the MRI gradient noise [9]. Sample videos are available at `http://sail.usc.edu/span/interspeech2008/index.php`.

5.3.3 Image analysis

In the midsagittal MR images the vocal tract contours were traced using the procedure described in [7]. Figure 5.2(a) shows an example midsagittal MR image of the fricative /s/ during its maximum constriction. Here the red lines delineate the outline of the vocal tract, and the green lines are the aperture lines which were computed using the methods described in [3].

In order to quantify the spatio-temporal shaping characteristics of the speech sounds of interest appropriate geometrical features have to be derived from each image. For the fricative sounds /s/ and /S/ one candidate feature is the midsagittal aperture at, and posterior to, the critical vocal tract constriction, which in these cases is formed using the tongue tip and the alveolar ridge/front palate. We hence proceed by defining a region



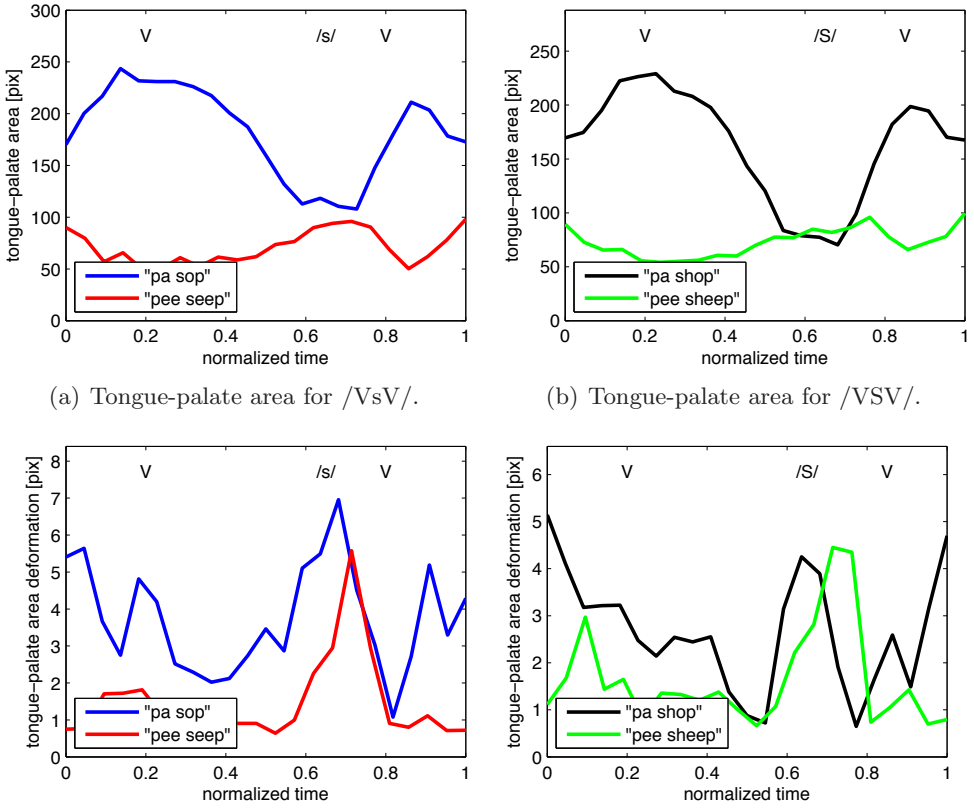
(a) Midsagittal RT-MRI image with vocal tract contours (red), aperture lines (green), and tongue-palate region boundaries (blue).

(b) Aperture function (green), tongue-palate region boundaries (blue), mean aperture in tongue-palate region (magenta), and aperture deformation area (black).

Figure 5.2: Midsagittal sample image and geometrical features during the fricative production in "pa seep."

of interest in the midsagittal profile of vocal tract bordered by the minimum opening between tongue and hard palate on the left (left blue line in Figure 5.2(a)) and a vertical line dropping from the hinge point of the velum (right blue line in Figure 5.2(a)). The selection of these boundaries is motivated by the relatively reliable detection of these anatomical landmarks. Figure 5.3(a) shows the time evolution of the size of this tongue-palate area for single tokens of the utterances "pee seep" and "pa sop." During the interval of the fricative production, which was identified for all tokens using the spectrogram of the synchronized audio recording, we observe a local minimum in the time function for the /a_a/ context and a local maximum for the /i_i/ context. The same holds for the /S/ sound as shown in Figure 5.3(b).

However, in order to better discriminate between the shaping difference of /s/ and /S/ it is also desired to devise a shape feature that is largely independent of the morphology of the subject's hard palate. To be more specific, we would like a measure of how parallel the palate and the tongue contours are in the region of interest in order to be able to deduce information on the nature of the airway channel. We hence propose the use of



(a) Tongue-palate area for /VsV/.

(b) Tongue-palate area for /VSV/.

(c) Tongue-palate area deformation for /VsV/.

(d) Tongue-palate area deformation for /VSV/.

Figure 5.3: Midsagittal features sample time functions.

a tongue-palate area *deformation* measure, which is illustrated in Figure 5.2(b). Here, the aperture function is shown (green) in addition to the location of the boundaries of the tongue-palate area of interest (vertical blue lines). As the deformation measure we use the variation (black shaded area) of the aperture about its mean value (magenta line) in the region of interest. This scalar *deformation* value will be near zero if the tongue is largely parallel to the hard palate, irrespective of the actual shape, and it will be large for non-parallel configurations. Figures 5.3(c) and 5.3(d) show sample time functions for the *deformation* feature for /s/ and /S/, respectively, for the /a_a/ and /i_i/ context, and we observe a local non-parallelity maximum during the interval of the

fricative production. We assume that at the maximum point the constriction target was achieved the preparation of the production of the following vowel will set it.

Our subsequent investigations will utilize the averaged maximum *deformation* over all tokens of a particular type, as well as the averaged area size at the time of maximum *deformation*. However, at this point we also want to point out a limitation of the proposed feature extraction process. As can be seen in Figure 5.1 the region of the hard palate is generally subject to rather larger image noise. This likely due to the fact that it consists of bone which is covered with a rather thin layer of soft-tissue. Since bone has a very low hydrogen content it produces a weak signal to noise ratio leading to faint MRI contrast. Hence the hard palate contour location, if left unconstrained during the automatic tracing process, can be adversely affected more than other sections of the vocal tract contours.

5.3.4 Coronal plane images

For the coronal images, the tongue contour was traced using a semi-automatic method [3]. Contours were initialized and corrected manually. Figure 5.4 shows an example coronal image with the traced tongue contour (red).

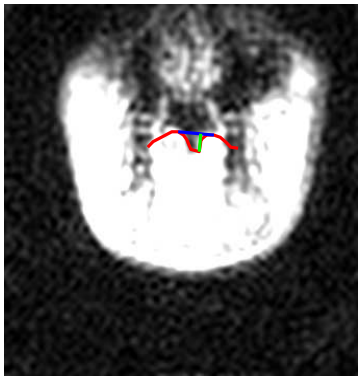


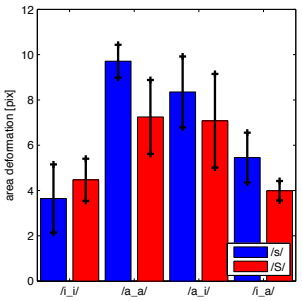
Figure 5.4: Coronal sample image, tongue contour (red), groove tangent (blue), grove depth feature (green) during the fricative production in “pa seep.”

As the analysis feature of interest we chose the coronal tongue groove depth. It is derived from the coronal tongue surface contour by finding the tangent (blue) onto the surface which forms the triangle with a maximum height (green). This is accomplished through an exhaustive search over all triangle combination of contour points for a given frame.

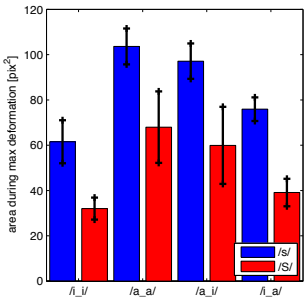
It should be noted that while the choice of the scan plane here was motivated to capture the nature of shaping in the region behind tongue front constriction (based on [43]), since it was anchored to an anatomical landmark, the images are not expected to align to any key shaping landmark such as maximal grooving for /s/ or doming for /S/, but instead provide an indication of the general shaping in that region.

5.4 Results

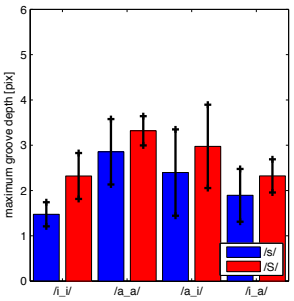
Using the measures discussed above, clear systematic differences between the two English fricatives /s/ and /S/ can be identified. Figures 5.5, 5.6, and 5.7 show the measurements for subjects A1, A2, and S1.



(a) Maximum tongue-palate area deformation.

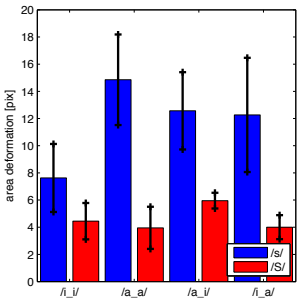


(b) Tongue-palate area at time of maximum deformation.

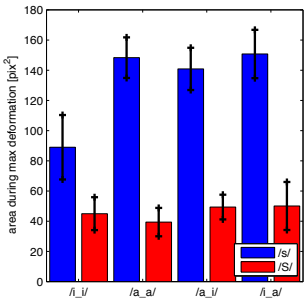


(c) Maximum tongue groove depth.

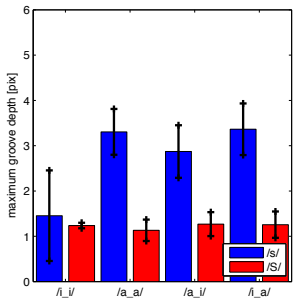
Figure 5.5: Subject A1 results.



(a) Maximum tongue-palate area deformation.

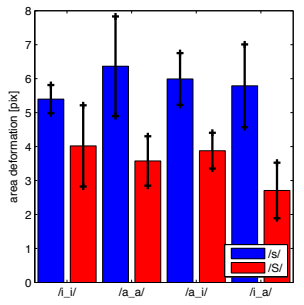


(b) Tongue-palate area at time of maximum deformation.

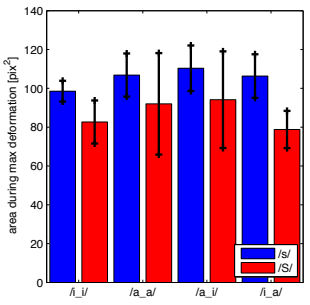


(c) Maximum tongue groove depth.

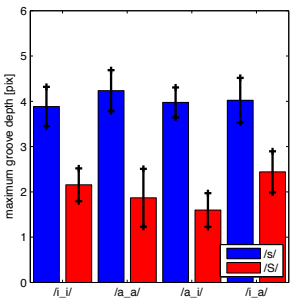
Figure 5.6: Subject A2 results.



(a) Maximum tongue-palate area deformation.



(b) Tongue-palate area at time of maximum deformation.



(c) Maximum tongue groove depth.

Figure 5.7: Subject S1 results.

The tongue-palate area at the time of maximum constriction shows robust differences between /s/ and /S/ for all three subjects. The fricative /s/ has a greater tongue-palate area at time of maximum constriction across all vowel contexts.

The differences in the measures were most robust for speaker A2. For tongue-palate area deformation and for tongue-palate area, clear differences are seen for the two fricatives. The maximum tongue-palate area deformation for speaker A2 was much greater for /s/ than it is for /S/ across all vowel contexts. The groove depth measurement for the coronal slice shown in Figure 5.6(c) also yields robust measurements. The grooving for /s/ was consistently seen in all subjects; the groove depth measured from the coronal slice was significantly greater under /a_a/ context than the /i_i/ context, with the other two

vocalic conditions considered showing intermediate values. The coronal plane choice for /S/ was further back in the oral cavity to capture the doming; rather, it provided a slice through the posterior tongue region that shows a cupping formation to support a raised doming in the anterior region [43]. In fact, the derived measures for /S/ do not show any significant variations indicating that this tongue back is not directly manipulated in the constriction formation.

The derived measures proposed in this paper are thus able to robustly account for the differences in articulation between /s/ and /S/.

5.5 Discussion

The measurements of MR images of the vocal tract discussed here provide useful techniques for studying the differences between the two English fricatives /s/ and /S/. The measurements of area and variance (midsagittal) and groove depth (coronal) are able to provide a way to distinguish between the two fricatives, which is not always a transparent and simple task.

Deriving concrete measurements of the articulatory properties of fricatives is necessary for further studies examining the shaping properties of fricatives. The real-time MRI techniques described here provide for fruitful analyses of the co-articulation effects between vowels and consonants. Examining the concavity of the fricatives with respect to the convexity of /i/ and the flatness of /a/ allows for several linguistic hypotheses to be addressed. The level of explicit control of tongue shape during fricative production is one such research question that will be considered using the data set discussed above.

Another possible avenue of research involves examining the variability across speakers. In our data, A1 uses a different part of the tongue tip to produce the sibilant fricatives than the other two subjects. Palatal morphology or other physiological differences could

play a role. The MR data and measures developed here provide for clear ways of answering this question.

Chapter 6

Statistical modeling of RT-MRI articulatory speech data

6.1 Abstract

This chapter investigates different statistical modeling frameworks for articulatory speech data obtained using RT-MRI. To quantitatively capture the spatio-temporal shaping process of the human vocal tract during speech production a multi-dimensional stream of direct image features is extracted automatically from the MRI recordings. The features are closely related, though not identical, to the tract variables commonly defined in the articulatory phonology theory. The modeling of the shaping process aims at decomposing the articulatory data streams into primitives by segmentation. A variety of approaches are investigated for carrying out the segmentation task including vector quantizer (VQ), Gaussian mixture model (GMM), hidden Markov model (HMM), and a CHMM. We evaluate the performance of the different segmentation schemes qualitatively with the help of a well understood data set which was used in an earlier study of inter-articulatory timing phenomena of American English nasal sounds.

6.2 Introduction

The recent technological advances in RT-MRI allow the speech researcher access to large quantities of rich articulatory data of running speech [6]. As opposed to previously available speech production data from EMA, which provides spatially sparse point tracking, and ultrasound, which is confined to capturing the tongue shape, RT-MRI captures the air-tissue boundaries along the entire vocal tract from the glottis to the lips. RT-MRI data hence appear to be a good basis for studying the vocal tract shaping process in a holistic way, i.e., they allow the investigation of individual articulators while simultaneously taking into account the effects of inter-articulatory coupling. However, the identification of shaping primitives from RT-MRI data (or from any other articulatory data) is not trivial, due to the data's high dimensionality, the complexity of the deformation space of the vocal tract, and the inter and intra subject variability in articulation.

In this article we will address the problem of identifying articulatory gestures from streams of RT-MRI image sequences. According to the theory of articulatory phonology [11], a gesture is a goal directed action of constriction forming by a vocal tract articulator. This process is modeled using the response of a second order linear system to a constriction target input step function. An articulator may be used to execute a sequence of consecutive gestures which leads to *temporal* gestural overlap. The gestures are quantified using tract variables, and it is important to realize that the mechanical coupling, due to the anatomical constraints, may produce *spatially* correlated measurement noise across different tract variables. So, the recognition of gestures from articulatory data must undo or at least take into account this spatio-temporal mixing.

For example, we can consider the lip aperture (LA) and tongue-tip constriction degree (TTCD) time series for the token /pay nova s/ as segmented from the carrier *"Type pay nova slowly."* (Fig. 6.1). Here, we have manually marked the critical constriction forming

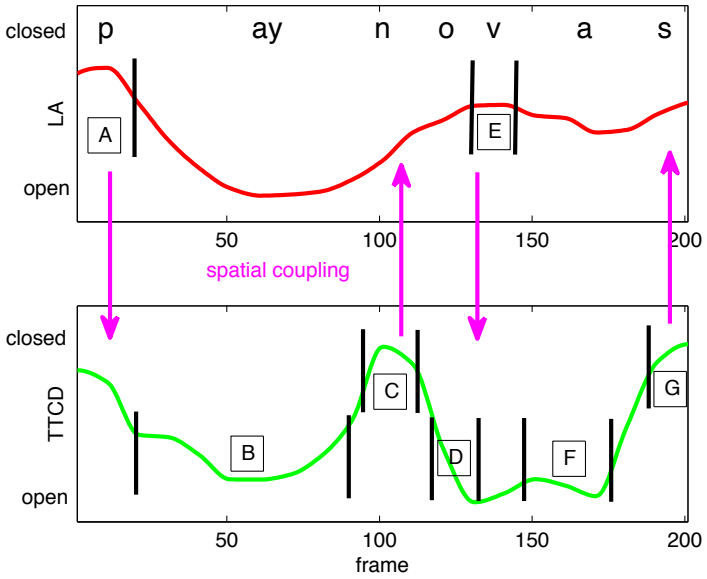


Figure 6.1: Lip aperture (LA) and tongue tip constriction degree (TTCD) time series for the utterance /pay nova s/ as derived from RT-MRI data (details given below).

processes. The segment labeled "A" of the LA trace corresponds to the bilabial closure for the formation of the /p/. Note that TTCD is also relatively constricted during this interval, due to articulatory coupling: the jaw contributes to lip closure, and brings the tongue tip towards the palate as a side-consequence. The purple arrow pointing down is meant to represent the direction of this coupling effect – from a phonologically controlled gesture to a passive coupling consequence. This is followed by the TTCD segment "B" for the formation of the diphthong /ay/. The diphthong is made using tongue body gestures which couple into the TTCD measurements. The subsequent tongue tip closure at the alveolar ridge in segment "C" is critical for the formation of the nasal, and we can identify a subtle effect on the LA trace due to the spatial coupling of the lips and the tongue via the jaw. As the tongue body is then used to produce the /o/ in segment "D", the lips move closer for the labiodental /v/ in segment "E," which again has an effect on the TTCD through spatial coupling. Finally, the production of the vowel /a/ ("F") with

the tongue body is followed by a period of narrow TTCD for the sibilant /s/ in segment "G."

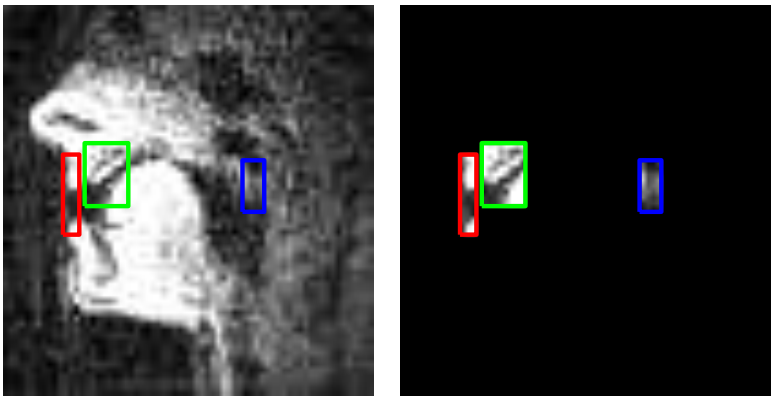
Previously, a variety of heuristic approaches have been pursued to model the shaping kinematics, such as the decomposition of individual EMA-traces into strokes [31], though with mixed results. In this paper we explore the use of a dynamical Bayesian network [42] to model the articulatory multi-stream data in a machine learning framework. Hereby, the joint modeling of different regions of the vocal tract is critical for coping with inter-articulator coupling, and the statistical processing will ensure a degree of robustness against intra subject variability.

This article is organized as follows. In Section 6.3 we will propose a simple yet robust way to obtain shaping information from the midsagittal MR images which aims at providing measurements closely related to the tract variables. Given a low-order parametric representation of the vocal tract shape we will, in Section 6.4, attempt a segmentation of image feature time series with VQ, GMM, uncoupled HMM, and a coupled HMM CHMM. The CHMM network is versatile, and it is particularly attractive since it is capable of handling asynchrony between data streams [46]. Finally, in Sections 6.5 and 6.6 we will discuss the results and draw conclusions.

6.3 Data preparation and parameterization

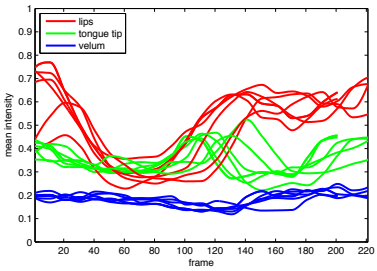
The data corpus for this case study consisted of two types of utterances produced by a female native American English speaker, namely *"Type pay nova slowly."* and *"Type pain over slowly."* The recordings were made using the scan protocol described in [45]. Seven realizations of each type, extracted from the carrier phrase, yielded the tokens /pay nova s/ and /pain over s/ used for our analysis. The starting frame was identified by the bilabial closure for /p/, and the end frame was chosen based on the narrow tongue

tip constriction at the alveolar ridge for /s/. The token duration was on the order of 1 second, and our MRI frame rate is approximately 22 frames per second. No timing normalization was carried out. A sample midsagittal MR image is shown in Fig. 6.2(a).

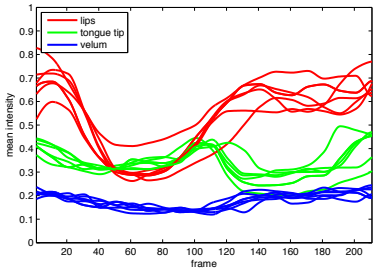


(a) Sample image.

(b) Masked image.



(c) Feature time series for 7 realizations of /pay nova s/.



(d) Feature time series for 7 realizations of /pain over s/.

Figure 6.2: Sample image and direct image feature time series.

The robust automatic extraction of the vocal tract shape in terms of its air tissue boundaries from the midsagittal MRI is not straightforward and is still considered to be an active domain of research [7, 20]. A versatile yet compact shape representation and parameterization, which would be beneficial for speech modeling purposes such as recognition, inversion, or synthesis, is not easy to obtain. Previous work in this domain includes the principal components based shape model used in [2, 39] or the constriction based vocal tract model implied by articulatory phonology [11]. Deriving such constriction measurements from image sequences can increase uncertainty of the data used for

modeling. Given the complex geometry of the vocal tract using a region based description of constriction events, rather than pinpointing a specific constriction location or its degree, appears to be a more robust choice. We focus on such a parameterization of the image sequences directly so as to capture the constriction events implicitly but robustly.

In this study we confine ourselves to investigating the articulatory processes involving the lips, the tongue tip, and the velum, and we select correspondingly in each image rectangular regions of interest as shown in Fig. 6.2(a) (shown as red, green, and blue box, respectively). The location of the regions is considered fixed, although this choice can also be dictated in a data driven way based on the region statistics such as the local image intensity correlation properties [33]. We can assume negligible head motion occurred during the experiment since the subjects head was well immobilized.

We then mask out the rest of MR image as shown in Fig. 6.2(b) and compute for each frame the average image intensity in each of the regions. The time series of these image intensity features are shown in Fig. 6.2(c) and 6.2(d) for all 7 realizations of /pay nova/ and /pain over/, respectively, and they have been ten-fold interpolated. The time series have a straightforward intuitive interpretation, since constriction forming events correspond to increasing the average image intensity because tissue moves into the particular region of interest. Conversely, a constriction release leads to a drop of average intensity over time since tissue moves out of the affected region. Hence the features closely resemble the constriction degree tract variables defined in articulatory phonology. Further, this representation can inherently capture the linguistically meaningful events in the presence of production variability, including due to inter-speaker morphological differences.

The two utterances were chosen because they differ minimally in the syllable position of the nasal, which is in coda position for /pain over/ and in onset position for /pay nova/. Previous studies [12, 4] have shown that systematic relative timing differences

exist for the tongue tip closure gesture and the velum opening gesture during the nasal production depending on its position in the utterance, and we will hence use this data set as a test case for our modeling framework.

6.4 Data modeling

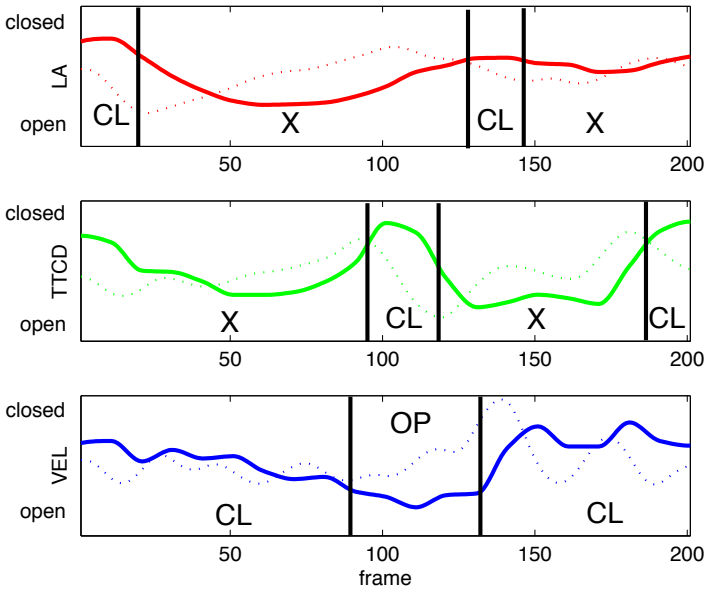


Figure 6.3: Lip aperture (LA), tongue tip constriction degree (TTCD), and velum aperture (VEL) for the utterance /pay nova s/ with gestural transcription. Solid line - feature time series, dashed line - first derivative.

Due to the limited number of training realizations in the data set considered, we will confine ourselves to detecting the gross shaping phenomena, i.e., the closure events "A," "C," "E," and "G" in Fig. 6.1. A simplified gestural transcription is shown in Fig. 6.3, where "OP" means open, "CL" means closed, and "X" means irrelevant state. The challenge for the segmentation algorithm will be to not give a false "CL" detection result at the very end of the LA trace (solid red), since that maximum is due to coupling from the TTCD (solid green). Equivalently, we would like no false "CL" alarm in the

beginning of the TTCD trace, since that maximum is due to spatial coupling with the LA trace. Both requirements are difficult to achieve robustly by a simple quantization of the time series. As noted earlier, the image sequences of vocal tract contours reflect a fairly complex dynamic geometry, and simple rule-driven ways of robustly identifying minimum constriction location/degree are difficult to implement, even with region based parameterization.

Generally, the time series data are quite noisy, and their first derivatives even more so (dotted lines in Fig. 6.3), especially for the velum (blue curves) due to the low image contrast in the pharyngeal region. So, rules such as through simple thresholding to find inflection points often do not yield reliable results. Hence, statistically capturing the time series behavior directly appears as a reasonable approach to pursue.

In the following we will augment the feature streams by their first derivatives, and attempt the modeling using VQ, GMM, HMM, and CHMM systems. These methods were chosen for a variety of reasons. The VQ is the most straightforward way to implement a simple instantaneous, i.e., time independent, thresholding mechanism for the individual 2-dimensional augmented feature data streams. The quantization levels can be found robustly using the well known k-means procedure, which, given the number of quantization levels, is otherwise parameter free. A manual transcription of all 14 data tokens as shown in Fig. 6.3 was produced, and it was used for the training of all of the methods. For the VQ, two centers were allocated corresponding to the two class labels. It should be noted that a VQ could also be implemented on the joint feature streams of all measurements, though we chose to keep the streams separate to allow "fair" comparisons of the VQ, GMM, and HMM methods.

The GMM can be considered a more sophisticated statistical way to achieve an instantaneous quantization, and it affords soft output values. However, in our case we

implemented subsequent hard clipping and thereby lose this advantage, but we included the GMM approach since it is often used in practice, and it can provide initialization parameters for the subsequent HMM systems. Just as the HMM and CHMM, the GMM is trained using the expectation maximization (EM) algorithm, which for all applications in this study was employed with a convergence threshold of 10^{-5}. The GMMs were implemented using the MATLAB Netlab toolbox which is a component of the BNT toolbox [42]. The models were initialized using k-means, and they had a full covariance matrix.

The HMM is a step up from the GMM in terms of modeling power and system complexity. It can be thought of as a time-dependent quantizer, and this method was chosen to address the temporal gestural overlap within a tract variable feature time series. Three individual HMMs were used for the LA, TTCD, and velic aperture (VEL) data. The HMMs were implemented using the MATLAB HMM toolbox which is also included in the BNT package. The hidden nodes had two states corresponding to the two segmentation labels used for each tract variable. Using the transcriptions, we initialized the observation models as bi-variate Gaussians with full covariance matrices, as well as the state priors and the ergodic state transition model.

The CHMM is the most complex system that we tested for this study, and it allows spatio-temporal modeling of the combined time series data. The model had three chains corresponding to the LA, TTCD, and VEL features (see Fig. 6.4), and it was implemented using the MATLAB BNT toolbox. The three hidden nodes had two states each, and the observations were bi-variate Gaussians. The CHMM parameters were initialized using the previously trained uncoupled HMMs.

We carried out the segmentation of our 14 observed articulatory traces using leave-one-out cross-validation, and we present in Table 6.1 some typical results. The graphs in

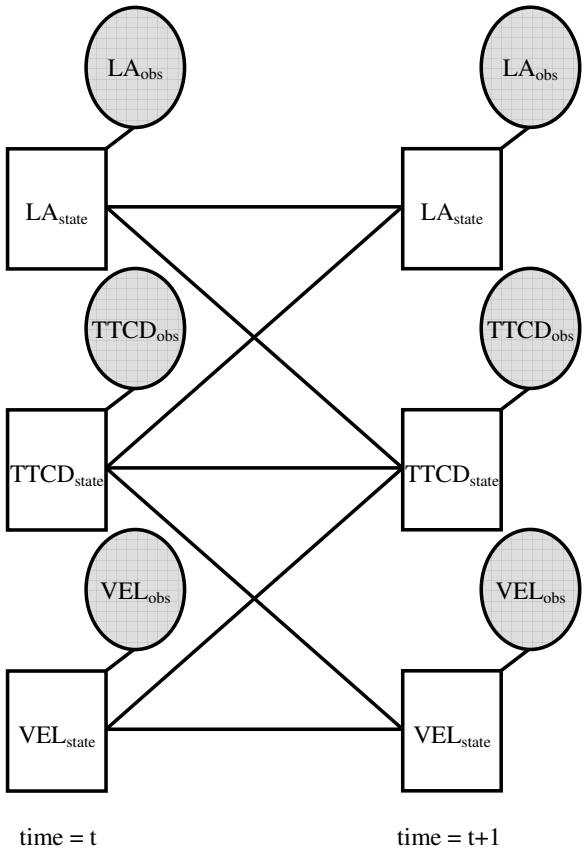
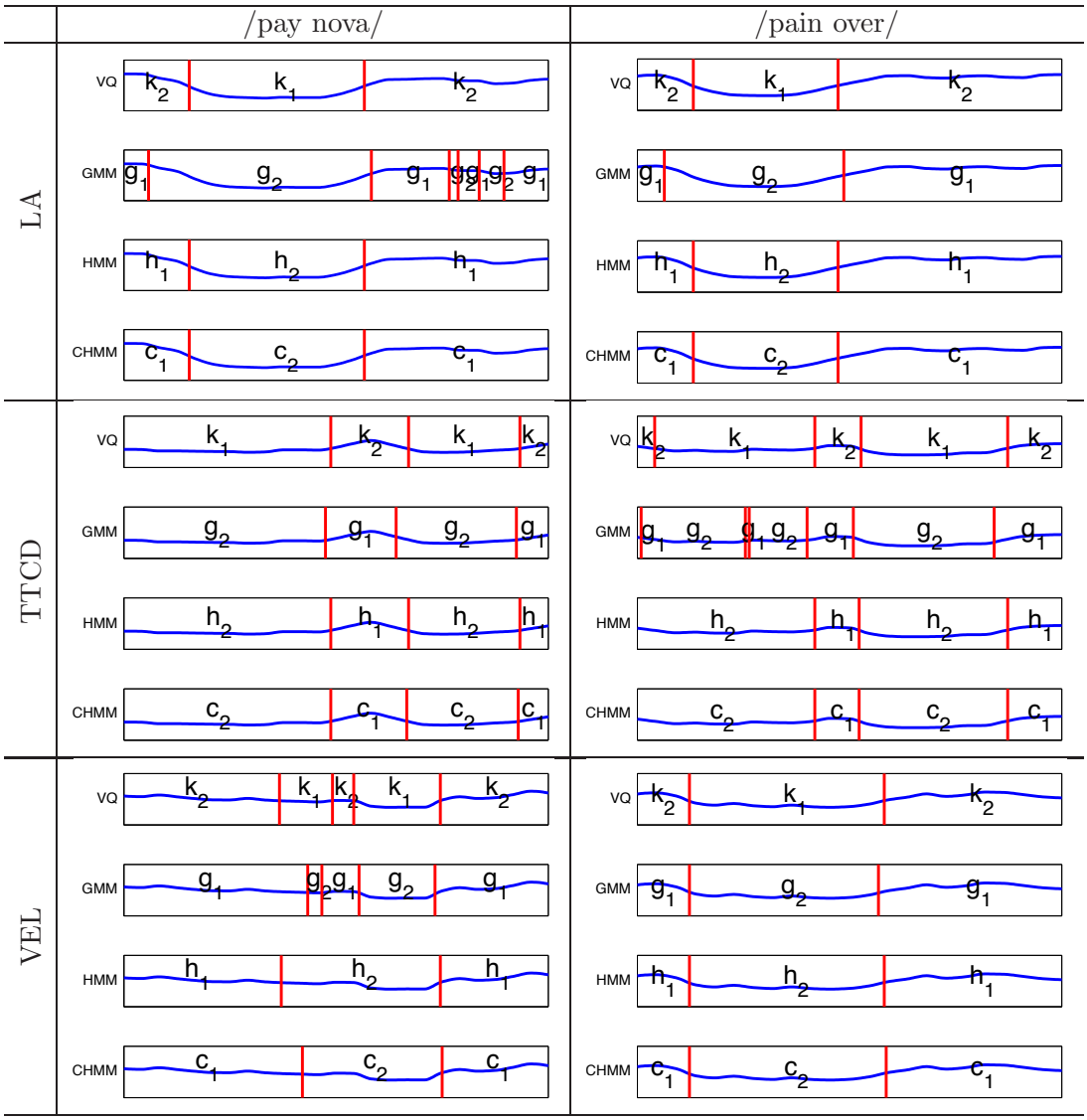


Figure 6.4: 3-chain CHMM layout (squares - hidden discrete nodes, shaded circles - continuous observations).

the left column correspond to a realization of /pay nova/ while the right column come from a realization of /pain over/. The top row shows the segmentation results for the LA trace, the middle row for TTCD, and the bottom row for VEL. The 4 plots in each row show the tract variable trace versus time (blue) and the segment boundaries (red vertical bars) as found by the VQ, GMM, HMM, and CHMM methods (top to bottom). The segments are labeled $k_{1,2}$ for VQ, $g_{1,2}$ for GMM, $h_{1,2}$ for HMM, and $c_{1,2}$ for CHMM.

Table 6.1: Sample segmentation results.



6.5 Discussion

In general we observed that the VQ and the GMM methods produced more spurious transitions, as shown for LA and VEL segmentation for /pay nova/, and TTCD segmentation for /pain over/ in Table 6.1. Generally, the HMM and the CHMM produce comparable and more consistent results.

With respect to the HMM and CHMM method, we found that both of them consistently labeled the initial bilabial closure segment in the LA trace. They also found the onset of the labiodental segment, but they repeatedly failed to identify its correct ending. However, both methods managed to avoid giving a false closure segment in the beginning of the TTCD trace. We found one realization of /pay nova/ for which the HMM, as opposed to all other methods, did not identify the VEL gesture at all.

Using the CHMM segmentation, we can now investigate the lag time difference between TTCD and VEL events for the formation of the nasal for the two types of tokens, i.e., we measure the time difference between the onset of the VEL opening (labeled c_2 in the bottom row, bottom graph of Table 6.1) and the onset of the TTCD closure (labeled c_1 in the center row, bottom graph). For the /pay nova/ tokens we obtain an average lag time of 96.8ms (σ=68ms), whereas for /pain over/ we obtain a lag of 279ms (σ=39.5ms). These results are encouraging since they are in accordance with previous findings [12, 4], and they seem to suggest that the proposed feature extraction procedure and the CHMM segmentation method appear to be robust and provide results that are consistent with our expectations.

In general we can suggest a number of ways to continue this study in order to improve the segmentation performance. On the one hand, one can certainly choose more complex models, e.g., higher-order mixtures for modeling the observations. And of course one can also scale up the entire procedure to include other image regions, leading to more chains in the CHMM. In any case, as more model parameters will have to be estimated a larger data corpus will be necessary. The possibility of collecting significant amounts of imaging data with RT-MRI holds promise in this regard.

6.6 Conclusions

We conclude from our study that the proposed method of image feature extraction has merit, and that the CHMM framework is a promising candidate for the discovery of articulatory primitives from RT-MRI data.

On a wider scope, this study indicates that if we combine (a) an explicit multistream transcription (gestures) with (b) appropriate techniques for extraction of articulatory time functions from RT-MRI data and with (c) the appropriate statistical models, we are well positioned to derive phonological information automatically from a rich set of articulatory data.

Chapter 7

Conclusions

7.1 Contributions

In summary, the core contributions of this thesis are (1) the design of a synchronized audio recording system, (2) the development of a model-based MRI noise cancellation algorithm, (3) the development of an upper-airway image-specific edge detection algorithm, (4) the pioneering use of RT-MRI data in speech production studies of soprano singing and fricative sounds, and (5) the introduction of a spatio-temporal statistical multistream modeling framework for RT-MRI data.

7.2 Future directions

Data is integral to advancing speech communication research, and vocal tract shaping information provides a crucial piece of the puzzle. There is an obvious need to gather and integrate multiple, disparate sources of information toward getting a more complete picture of the underlying processes.

The problem is highly challenging from both the technological as well as the theoretical perspective. In the future there is tremendous potential for applications including machine recognition, coding, and synthesis of speech.

Above all, acquiring, interpreting and utilizing speech production information is an ongoing interdisciplinary scientific endeavor.

Glossary

ADC analog to digital converter. 9, 10

CHMM coupled hidden Markov model. xi, 85, 88, 92–97

EMA electromagnetometry. 2, 73, 86

FIR finite impulse response. 12, 13, 15

FOV field of view. 19, 29–32

FPGA field programmable gate array. 5, 9

GMM Gaussian mixture model. 85, 88, 92–95

HMM hidden Markov model. 85, 88, 92–96

LA lip aperture. 86, 87, 91–96

MR magnetic resonance. x, 1–3, 6, 7, 18, 19, 21–23, 28–30, 41, 42, 45, 47, 52–54, 56–60, 71, 74, 77, 83, 88, 90

MRI magnetic resonance imaging. ix–xi, 1–3, 5–13, 15, 19, 23, 60, 61, 63, 64, 70–72, 74–77, 80, 85, 86, 96–98

NLMS normalized least mean square. 10, 12, 13

RF radio frequency. 8–10, 19

RSS root sum square. 42

RT real time. ix–xi, 1–3, 18, 19, 61, 63, 64, 70–72, 74, 85, 86, 96–98

SNR signal to noise ratio. 15, 16

TE echo time. 11

TR repetition time. 9, 10, 12, 19

TTCD tongue-tip constriction degree. 86–88, 91–96

VEL velic aperture. 93–96

VQ vector quantizer. 85, 88, 92, 94, 95

References

[1] P. Badin. Acoustics of voiceless fricatives: Production theory and data. *Speech Technol. Lett.*, pages 45–52, Apr.-Sep. 1989.

[2] P. Badin, G. Bailly, L. Revéret, M. Baciu, C. Segebarth, and C. Savariaux. Three-dimensional linear articulatory modeling of tongue, lips and face based on MR and video images. *Journal of Phonetics*, 30:533–553, 2002.

[3] E. Bresch, J. Adams, A. Pouzet, S. Lee, D. Byrd, and S. Narayanan. Semi-automatic processing of real-time MR image sequences for speech production studies. In *Proc. 7th International Seminar on Speech Production*, Ubatuba, Brazil, December 2006.

[4] E. Bresch, L. Goldstein, and S. Narayanan. An analysis-by-synthesis approach to modeling real-time mri articulatory data using the task dynamic application framework. *157th Meeting of the Acoustical Society of America*, May 2009.

[5] E. Bresch, N. Katsamanis, L. Goldstein, and S. Narayanan. Statistical multi-stream modeling of real-time MRI articulatory speech data. *Interspeech*, 2010.

[6] E. Bresch, Y.-C. Kim, K. Nayak, D. Byrd, and S. Narayanan. Seeing speech: capturing vocal tract shaping using real-time magnetic resonance imaging. *IEEE Signal Processing Mag.*, 25(3):123–132, 2008.

[7] E. Bresch and S. Narayanan. Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images. *IEEE Trans. Med. Imag.*, 28(3):323–338, March 2009.

[8] E. Bresch and S. Narayanan. Real-time MRI investigation of resonance tuning in soprano singing. *J. Acoust. Soc. Am.*, 128(5):EL335 – EL341, November 2010.

[9] E. Bresch, J. Nielsen, K. Nayak, and S. Narayanan. Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans. *J. Acoust. Soc. Am.*, 120(4):1791–1794, October 2006.

[10] E. Bresch, D. Riggs, L. Goldstein, D. Byrd, S. Lee, and S. Narayanan. An analysis of vocal tract shaping in english sibilant fricatives using real-time magnetic resonance imaging. *Interspeech*, 2008.

[11] C. Browman and L. Goldstein. Towards an articulatory phonology. *Phonology Yearbook*, 3:219–252, 1986.

[12] D. Byrd, S. Tobin, E. Bresch, and S. Narayanan. Timing effects of syllable structure and stress on nasals: a real-time MRI examination. *Journal of Phonetics*, 37:97–110, 2009.

[13] G. Carlsson and J. Sundberg. Formant frequency tuning in singing. *Journal of Voice*, 6:256–260, 1992.

[14] G. Charpiat, R. Keriven, J.P. Pons, and O. Faugeras. Designing spatially coherent minimizing flows for variational problems based on active contours. In *Proc. IEEE Tenth International Conference on Computer Vision*, 2005.

[15] F. L. Chu and C. F. Huang. On the calculation of the Fourier transform of a polygonal shape function. *Journal of Physics A: Mathematical and General*, 22(9):L671–L672, 1989.

[16] I. Eckstein, J.-P. Pons, Y. Tong, C.-C.J.Kuo, and M. Desbrun. Generalized surface flows for mesh processing. In *SGP (Eurographics Symposium on Geometry Processing)*, pages 183–192, Barcelona, Spain, 2007.

[17] O. Engwall and P. Badin. An MRI study of Swedish fricatives: coarticulatory effects. In *5th Speech Production Seminar*, 2000.

[18] M.A.T. Figuereido, J.M.N. Leitao, and A.K. Jain. Unsupervised contour representation and estimation using b-splines and a minimum descrition length criterion. *IEEE Trans. Image Processing*, 9(6):1075–1087, 2000.

[19] J. Fontecave and F. Berthommier. Semi-automatic extraction of vocal tract movements from cineradiographic data. In *Proceedings of the Ninth International Conference on Spoken Language Processing*, pages 569–572, Pittsburgh, PA, 2006.

[20] J. Fontecave and F. Berthommier. A semi-automatic method for extracting vocal tract movements from x-ray films. *Speech Communication*, 51:97–115, 2008.

[21] J. Glover. Adaptive noise canceling applied to sinusoidal interferences. *IEEE Trans. Acoustics, Speech and Signal Processing*, 25(6):484–491, 1977.

[22] G. Golub and V. Pereyra. Separable nonlinear least squares: the variable projection method and its applications. In *Inverse Problems*, volume 19, pages R1–R26. Institute of Physics Publishing, 2003.

[23] S. Haykin. *Adaptive filter theory*. Prentice Hall, Upper Saddle River, NJ, 2001.

[24] G. Huang and C.C. Kuo. Wavelet descriptor of planar curves: Theory and application. *IEEE Trans. Image Processing*, 5(1):56–70, 1996.

[25] J. I. Jackson, C. H. Meyer, D. G. Nishimura, and A. Macovski. Selection of a convolution function for Fourier inversion using gridding. *IEEE Trans. Med. Imaging*, 10(3):473–478, September 1991.

[26] E. Joliveau, J. Smith, and J. Wolfe. Tuning of vocal tract resonance by sopranos. *Nature*, 427(116), 2004.

[27] E. Joliveau, J. Smith, and J. Wolfe. Vocal tract resonances in singing: The soprano voice. *J. Acoust. Soc. Am.*, 116:2434–2439, 2004.

[28] D. Jones. Normalized LMS. `http://cnx.rice.edu/content/m11915/latest/`, February 2006.

[29] Y. Kahana, A. Paritsky, A. Kots, and S. Mican. Recent advances in optical microphone technology. In *Proc. of the 32nd International Congress and Exposition on Noise Control Engineering*, 2003.

[30] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal on Computer Vision*, pages 321–331, 1988.

[31] T. Kato, S. Lee, and S. Narayanan. An analysis of articulatory-acoustic data based on articulatory strokes. *Proc. IEEE Int'l Conf. Acous., Speech, and Signal Processing*, pages 4493–4496, 2009.

[32] P. Ladefoged, J.F.K. Anthony, and C. Riley. Direct measurement of the vocal tract. *UCLA working papers in phonetics*, 19:4–13, 1971.

[33] A. Lammert, M. Proctor, and S. Narayanan. Data-driven analysis of realtime vocal tract MRI using correlated image regions. *Interspeech*, 2010.

[34] L. Lee, P. Fieguth, and L. Deng. A functional articulatory dynamic model for speech production. In *Proc. IEEE 2001 International Conference on Acoustics, Speech, and Signal Processing (ICASSP '01)*, volume 2, pages 797–800, 2001.

[35] S. Lee. *A study of vowel articulation in a perceptual space*. PhD thesis, University of Alabama at Birmingham, 1991.

[36] S. W. Lee and R. Mitra. Fourier transform of a polygonal shape function and its application in electromagnetics. *IEEE Trans. Antennas Propagat.*, AP-31(1):99–103, 1983.

[37] F. Li, J. Edwards, and M. Beckman. Spectral measures for sibilant fricatives of English, Japanese, and Mandarin Chinese. In *Proc. 16th International Congress of Phonetic Sciences*, pages 917–920, Saarbrücken, Germany, August 2007.

[38] C. Liu, R. Bammer, and M. Moseley. Parallel imaging reconstruction for arbitrary trajectories using k-space sparse matrices (kSPA). *Magn. Reson. Med.*, 56, 2007.

[39] S. Maeda. Compensatory articulation during speech: evidence from the analysis and synthesis of vocal tract shapes using an articulatory model. In W.J. Hardacastle and A. Marchal, editors, *Speech Production and Speech Modeling*. Kluwer Academic Publishers, 1990.

[40] T. McInerney and D. Terzopoulos. Deformable models in medical image analysis: A survey. *Medical Image Analysis*, 1(2):91–108, 1996.

[41] K. McInturff and P.S. Simon. The Fourier transform of linearly varying functions with polygonal support. *IEEE Trans. Antennas Propagat.*, 39(9), 1991.

[42] K. Murphy. The Bayes Net Toolbox for Matlab. *Computing Science and Statistics*, 33, 2001.

[43] S. Narayanan, A. Alwan, and K. Haker. An articulatory study of fricative consonants using magnetic resonance imaging. *J. Acoust. Soc. Am.*, 98(3):1325–1347, September 1995.

[44] S. Narayanan, A. Alwan, and Y. Song. New results in vowel production: MRI, EPG, and acoustic data. In *Proc. EuroSpeech*, volume 1, pages 1007–1010, Rhodes, Greece, September 1997.

[45] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd. An approach to real-time magnetic resonance imaging for speech production. *J. Acoust. Soc. Am.*, 115:1771–1776, 2004.

[46] A. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy. Dynamic bayesian networks for audio-visual speech recognition. *J. Applied Signal Processing*, 2002.

[47] M. NessAiver, M. Stone, V. Parthasarathy, Y. Kahana, and A. Paritsky. Recording high quality speech during tagged cine mri studies using a fiber optic microphone. *J. Magn. Reson. Imaging*, 23:92–97, 2006.

[48] J. Perkell, M. Cohen, M. Svirsky, M. Matthies, I. Garabieta, and M. Jackson. Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *J. Acoust. Soc. Am.*, 92(6):3078–1049, 1992.

[49] M. Proctor, C. Shadle, and K. Iskarous. An MRI study of vocalic context effects and lip rounding in the production of english sibilants. In *11th Australian International Conference on Speech Science and Technology*, pages 307–312, University of Auckland, New Zealand, December 2006.

[50] E. L. Saltzman and K. G. Munhall. A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1(4):333–382, 1989.

[51] J. M. Santos, G. A. Wright, and J. M. Pauly. Flexible real-time magnetic resonance imaging framework. In *Proc., IEEE EMBS, 26th Annual Meeting*, San Francisco, 2004.

[52] N. A. Schmid, Y. Bresler, and P. Moulin. Complexity regularized shape estimation from noisy Fourier data. In *Proc. IEEE 2002 International Conference on Image Processing*, pages 453–456, 2002.

[53] C. H. Shadle. The acoustics of fricative consonants. *Tech. Rep. 506*, pages 45–52, 1985.

[54] A. Soquet, V. Lecuit, T. Metens, and D. Demolin. Mid-sagittal cut to area function transformations: Direct measurements of mid-sagittal distance and area with MRI. *Speech Communication*, 36:169–180, 2002.

[55] M. Stone. Toward a model of three-dimensional tongue movement. *Journal of Phonetics*, 19:309–320, 1991.

[56] M. Stone, E. Davis, A. Douglas, M. NessAiver, R. Gullapalli, W. Levine, and A. Lundberg. Modeling tongue surface contours from cine-MRI images. *Journal of Speech, Language, and Hearing Research*, 44(5):1026–1040, 2001.

[57] M. Stone, E. Davis, A. Douglas, M. NessAiver, R. Gullapalli, W. Levine, and A. Lundberg. Modelling the motion of the internal tongue from tagged cine-mri images. *J. Acoust. Soc. Am.*, 109(6):2974–2982, June 2001.

[58] B. Story. Vowel acoustics for speaking and singing. *Acta Acustica united with Acustica*, 90:629–640, 2004.

[59] B. H. Story. Using imaging and modeling techniques to understand the relation between vocal tract shape to acoustic characteristics. In *Proc. Stockholm Music Acoustics Conference SMAC-03*, pages 435–438, 2003.

[60] J. Sundberg. The acoustics of the singing voice. *Sci Am.*, 236:82–91, 1977.

[61] J. Sundberg. Research on the singing voice in retrospect. *TMH-QPSR*, 45(1):11–22, 2003. KTH Stockholm.

[62] H. Takemoto, K. Honda, S. Masaki, Y. Shimeda, and I. Fujimoto. Measurement of temporal changes in vocal tract area function from 3d cine-mri data. *J. Acoust. Soc. Am.*, 119(2):1037–1049, 2006.

[63] I. Titze. A theoretical study of f0-f1 interaction with application to resonant speaking and singing voice. *Journal of Voice*, 18:292–298, 2004.

[64] D. Whalen, K. Iskarous, M. Tiede, and D. Ostry. The Haskins optically corrected ultrasound system (HOCUS). *Journal of Speech, Language, and Hearing Research*, 48(3):543–554, 2005.

[65] B. Widrow, K. Duvall, R. Gooch, and W. Newman. Signal cancellation phenomena in adaptive antennas: causes and cures. *IEEE Trans. Antennas Propagat.*, 30(3):469–478, 1982.

[66] Z. Zhang and C. Y. Espy-Wilson. A Vocal-tract model of American English /l/. *J. Acoust. Soc. Am.*, 115(3):1274–1280, March 2004.

Appendix

Fourier transform of a polygonal shape function and the vertex vector derivative

A simple polygon P with the counterclockwise oriented vertices $\mathbf{v}_i = [x_i, y_i], i = 0 \ldots (M-1)$ in the $x-y$ plane is shown in Fig. A.1. With

$$\mathbf{V} = [\mathbf{v}_0^T, \ldots, \mathbf{v}_{M-1}^T]^T \tag{A.1}$$

we can define the shape function

$$s(x, y, \mathbf{V}) = \begin{cases} 1 & \text{for } (x, y) \text{ in } P(\mathbf{V}) \\ 0 & \text{otherwise} \end{cases} \tag{A.2}$$

and the corresponding two-dimensional Fourier transform

$$\begin{aligned} S(k_x, k_y, \mathbf{V}) &= \mathcal{F}\{s(x, y, \mathbf{V})\} \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} s(x, y, \mathbf{V}) e^{-j2\pi(k_x x + k_y y)} dx dy \\ &= \iint_{(x,y)\text{inside}P(\mathbf{V})} e^{-j2\pi(k_x x + k_y y)} dx dy \end{aligned} \tag{A.3}$$

where x and y are the spatial coordinates, and k_x and k_y are the spatial frequency variables.

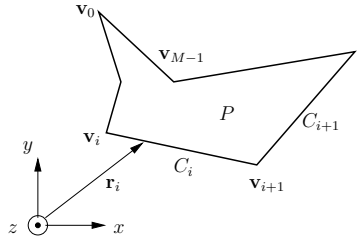


Figure A.1: A simple polygon P

Analytic expressions for the Fourier transform $S(k_x, k_y, \mathbf{V})$ have been derived using various methods in numerous articles such as [36], [15], and [41]. The latter contribution elegantly employs Stokes' Theorem and yields for $k_x = 0$ and $k_y = 0$

$$S(k_x, k_y, \mathbf{V}) = \frac{1}{2} \sum_{i=0}^{M-1} (x_i y_{i+1} - x_{i+1} y_i) \tag{A.4}$$

and otherwise

$$S(k_x, k_y, \mathbf{V}) = \frac{1}{j2\pi(k_x^2 + k_y^2)} \sum_{i=0}^{M-1} \left(\operatorname{sinc}\left((\mathbf{v}_{i+1} - \mathbf{v}_i) \cdot \begin{bmatrix} k_x \\ k_y \end{bmatrix} \right) e^{-j\pi(\mathbf{v}_{i+1} + \mathbf{v}_i) \cdot \begin{bmatrix} k_x \\ k_y \end{bmatrix}} (\mathbf{v}_{i+1} - \mathbf{v}_i) \cdot \begin{bmatrix} k_y \\ -k_x \end{bmatrix} \right) \tag{A.5}$$

where $\operatorname{sinc}(x) = \frac{\sin(\pi x)}{\pi x}$.

With the help of the chain rule we can now find an expression the derivative of $S(k_x, k_y, \mathbf{V})$ with respect to a vertex vector $\mathbf{v}_i$

$$\frac{\partial S(k_x, k_y, \mathbf{V})}{\partial \mathbf{v}_i} = \left[\frac{\partial S(k_x, k_y, \mathbf{V})}{\partial x_i}, \frac{\partial S(k_x, k_y, \mathbf{V})}{\partial y_i} \right] \tag{A.6}$$

and with

$$\mathbf{a}_i = \mathbf{v}_{i+1} - \mathbf{v}_i \tag{A.7}$$

$$\mathbf{b}_i = \mathbf{v}_{i+1} + \mathbf{v}_i \tag{A.8}$$

$$c_i = -k_x \operatorname{sinc}'\left(\mathbf{a}_i \cdot \begin{bmatrix} k_x \\ k_y \end{bmatrix} \right) e^{-j\pi \mathbf{b}_i \cdot \begin{bmatrix} k_x \\ k_y \end{bmatrix}} \left(\mathbf{a}_i \cdot \begin{bmatrix} k_y \\ -k_x \end{bmatrix} \right) \tag{A.9}$$

$$d_i = -j\pi k_x \operatorname{sinc}\left(\mathbf{a}_i \cdot \begin{bmatrix} k_x \\ k_y \end{bmatrix} \right) e^{-j\pi \mathbf{b}_i \cdot \begin{bmatrix} k_x \\ k_y \end{bmatrix}} \left(\mathbf{a}_i \cdot \begin{bmatrix} k_y \\ -k_x \end{bmatrix} \right) \tag{A.10}$$

$$e_i = -k_y \operatorname{sinc}\left(\mathbf{a}_i \cdot \begin{bmatrix} k_x \\ k_y \end{bmatrix} \right) e^{-j\pi \mathbf{b}_i \cdot \begin{bmatrix} k_x \\ k_y \end{bmatrix}} \tag{A.11}$$

where $\operatorname{sinc}'(x) = \frac{d \operatorname{sinc}(x)}{dx}$, we finally obtain

$$\frac{\partial S(k_x, k_y, \mathbf{V})}{\partial x_i} = \begin{cases} \frac{y_{i+1} - y_{i-1}}{2} & \text{for } k_x = 0, k_y = 0 \\ \frac{c_i + d_i + e_i - c_{i-1} + d_{i-1} - e_{i-1}}{j2\pi(k_x^2 + k_y^2)} & \text{otherwise} \end{cases} \tag{A.12}$$

The expression for $\frac{\partial S(k_x,k_y,\mathbf{V})}{\partial y_i}$ can be derived along the same lines and is omitted here.